

AD _____

Award Number: DAMD17-98-1-8119

TITLE: Genetic Damage Caused by ALU Repeats in Breast Cancer

PRINCIPAL INVESTIGATOR: Prescott L. Deininger, Ph.D.

CONTRACTING ORGANIZATION: Tulane University Medical Center
New Orleans, Louisiana 70112-2699

REPORT DATE: August 2001

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED	
	August 2001	Final (1 Aug 98 - 31 Jul 01)	
4. TITLE AND SUBTITLE		5. FUNDING NUMBERS	
Genetic Damage Caused by ALU Repeats in Breast Cancer		DAMD17-98-1-8119	
6. AUTHOR(S)		8. PERFORMING ORGANIZATION REPORT NUMBER	
Prescott L. Deininger, Ph.D.			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)		9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)	
Tulane University Medical Center New Orleans, Louisiana 70112-2699 E-Mail: pdeinin@tulane.edu		U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	
10. SPONSORING / MONITORING AGENCY REPORT NUMBER		11. SUPPLEMENTARY NOTES	
		Report contains color	
12a. DISTRIBUTION / AVAILABILITY STATEMENT		12b. DISTRIBUTION CODE	
Approved for Public Release; Distribution Unlimited			
13. ABSTRACT (Maximum 200 Words)			
<p>We have developed a series of allele-specific PCR amplification procedures that allow us to amplify the flanking sequences from the most recent subfamilies of Alu elements in the human genome. There are approximately 1000 elements amplified in these experiments, and we have developed several strategies for amplifying specific subsets of these elements. The goal is to identify subsets of elements that can be amplified and 'displayed' on a gel-based or subtractive method that will allow us to detect differences in these recent elements in breast tumor vs. normal tissue from a patient. This will allow us to detect either insertion of a new Alu element and assessment of the rate of gene damage from retrotransposition, as well as detect major sequence losses that encompass one of these elements.</p>			
14. SUBJECT TERMS		15. NUMBER OF PAGES	80
		16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

(4) Table of Contents

Front Cover	p. 1
Form 298	p. 2
Table of Contents	p. 3
Introduction	p. 4
Body	p. 4-7
Key Research Accomplishments	p. 8
Reportable Outcomes	p. 8
Conclusions	p. 9
References	p. 9
Appendices	p. 9 and 5 appended reprints

(5) Introduction:

This project was based on the hypothesis that early cellular transformation events involved in breast cancer formation might influence the amplification of human Alu repeats. Any increases in Alu amplification, might contribute to further destabilization of the human genome and inactivation of tumor suppressors that could contribute to the progression of breast cancer. At least in sporadic cases, Alu insertions have been shown to contribute to a number of cancers, including at least one case of breast cancer due to inactivation of BRCA2¹. We have previously shown that only a specific set of subfamilies of Alu elements are actively amplifying in the human genome^{2,3}. This project combines this information with an anchored PCR procedure we have developed to form displays of the most recently amplified Alu elements. We have demonstrated that this Allele-Specific Alu PCR (ASAP) will effectively display the members of the smallest of the recent Alu subfamilies as bands on an acrylamide gel (5). Our goal is to generalize these procedures to the larger subfamilies and explore various procedures to deal with the larger number of bands expected. We will then use these procedures to compare breast cancer and normal DNA from a number of individuals to determine whether there are new, tumor-specific Alu inserts. This will allow us to determine whether this form of genetic instability plays a role in human breast cancer.

Because of some difficulties with initial implementation of the ASAP assay, we also designed approaches to use an L1 retrotransposition reporter gene system (Moran) to study the specific influences on retrotransposition of genetic changes associated with tumorigenesis, as well as environmental influences that may contribute to breast cancer. This will allow. Because it is thought that Alu elements utilize the same retrotransposition machinery as L1, this system should allow an alternate assessment of the primary question of whether retroelement insertions are likely to contribute to breast cancer genomic instability.

(6) BODY

Original Goals:

First Six Months:

- Optimization of ASAP. Our primary goal will be to optimize the Allele-Specific PCR further. We will work to identify the very best PCR primers to allow the most effective allele-specific amplification of the Alu inserts and flanks. This will allow us to develop a procedure with both minimal steps and minimal background in the later experiments.
- No patient samples will be needed at this stage.

First Year:

- Optimization of Displays. We will utilize the ASAP procedure to generate test samples from all three relevant Alu subfamilies, which can then be utilized to improve the display procedures, in particular the subdivision with PCR into 16 subdivisions. We will begin to explore ways to utilize subtraction procedures on these samples.
- No patient samples will be needed at this stage

Second Year:

- Refinement of Subtraction Technology. Technical development will continue with refinement of the subtraction procedures and tests of the sensitivity of detection of bands and the ability to pool samples in the PCR reactions.
- Preliminary work on tumor samples. Work will begin with existing technology to carry out analysis on tumor samples. We expect to have carried out analysis of the first 10-20 samples in this year. We will use this experience to determine the best approach to generate data in a production mode. This will provide an initial feel for the level of diversity in the displays and a basic characterization of any diversity to determine whether it is caused by insertions. Any evidence of other forms of genomic instability influencing the assay will be assessed at this point and procedures optimized to compensate.

Third Year:

- Completion of Tumor Samples. During the previous year, we expect to have optimized the ASAP procedures and their display completely. This will allow us to have determined the most effective approach for analysis of large numbers of samples. We will utilize this year solely to generate data on as many tumors as possible. We will focus our efforts initially on late stage tumors, but will move progressively towards earlier stage tumors, particularly if we detect extensive Alu amplification at late stages.
- We expect to complete 100 samples by the end of the third year. It is our hope that the subtraction of pooled samples will increase the data flow and we can carry out experiments on enough samples to be able to analyze subgroups based on tumor stage, ethnic origin of tumor or other correlations with clinical features or treatment.

By the Second year it became clear that there were more technical difficulties getting the displays fully optimized and implementable on a large number of samples and our goals had to be scaled back to a more pilot level. In addition, last year we reported in our progress report an alternative approach to address the critical issue of whether retrotransposition played a critical role in breast cancer progression. The approach was to use a reporter system for L1 retrotransposition and test whether genetic alterations associated with tumorigenesis altered retrotransposition rates.

Accomplishments of the three year period:

(This includes a summary of the first two year's work, although without the detail placed in those reports).

During the first two years we explored a wide range of approaches for optimizing displays of the most recently inserted Alu inserts. Year 1 focused primarily on the PCR-based display itself, utilizing a number of variations to both increase the resolution of the technique, as well as ways to deal with the large numbers of elements in some of the more active subfamilies which gave rise to too many elements to allow our assay to work. We were successful at generating quality displays for the very smallest subfamilies of elements. We also had some success utilize various less frequent restriction digestions to allow us to display a limited subset of the more abundant subfamilies. Our biggest difficult at this point was to figure out how to display the 2000 Ya5 subfamily members (which are responsible for the majority of Alu inserts causing disease), without the massive number of bands obscuring the variant signals. We had

limited success with the use of PCR primers that added two bases to the end of the primer that went into the genomic flanking sequence to allow us to display one sixteenth of the group of bands at a time. Several primers gave use decent, although not crisp displays. I believe that our biggest problem with this approach was that some of the primers could sit down on sites in which the last two bases base-paired using non Watson-Crick pairing (i.e. G-T pairing), resulting in weaker bands that created background. In our efforts, although several primers worked pretty well, others worked very poorly. A number of variants (include perfect match, altering stringency, etc) did not improve these displays ultimately. Perhaps our biggest disappointment was that several attempts to utilize subtraction strategies to eliminate the common bands did not work at all. Our only observation was that the bands all got lighter, but even attempts to spike a unique band in the mix did not allow us to enrich the unique band. These studies may have been influenced by the presence of a small segment of common repetitive DNA sequence on the end of each fragment, and they may have also been made more difficult by the very high A+T content of the sequences adjacent to Alu elements.

As more human genomic sequence was made available in GENBANK, we were able to identify new subfamilies of Alu elements. More importantly, we found that some of the subfamilies showed very high levels of polymorphism in the human genome. Using a combination of bioinformatics with measurements of the polymorphism associated with these different subfamilies, we were able to determine the relative age and copy number of each of their subfamilies and provide estimates of their likelihood of current activity. Although these data did provide some new, smaller subfamilies that we could adapt to our display technique, by far the majority of Alu elements that had inserted recently to cause disease still remained as part of the larger Ya5 and Yb8 subfamilies. Thus, our original plan of displaying the majority of potential Alu inserts in tumor DNA was not going to work with this approach.

As we approached year 3, we also began to tackle some of the issues associated with adapting this technique to a number of tumor tissues to allow a reasonable sampling. If anything the tumor tissues were even more intractable, partly because the DNA was not always of as high a quality as the tissue culture DNA, and blood DNAs, that we were using in the pilot experiments. Furthermore, our display would be seriously handicapped by any heterogeneity in the tumor tissue that might weaken the signals, while not lessening the background. Therefore, although we worked out the ability to display distinct subsets of the recent Alu inserts, we were never able to adapt the technique to be able to display a significant portion of these inserts in a manner which convinced us that we would be able to see any significant portion of new inserts. Given that new inserts may have been as low as one in 100 tumors, we began to explore alternative approaches for addressing the potential role of retrotransposition in breast cancers.

Although the ideal was to look at authentic tumor tissues and look for authentic Alu inserts, we would obtain a pretty good picture of the relative impact by using a reporter system introduced into tumor cells and measuring the rate of retrotransposition of the reporter system in normal versus transformed cells. The development of an L1 element that activated a neomycin selection cassette upon retrotransposition, provided a potential method to quantify L1 retrotransposition rates in tumors⁴. Furthermore, as most of us believe that Alu retrotransposes with the L1 machinery, using the L1 system should provide insight into both L1 and Alu rates.

Our initial experiments using p53 transformation as a model were very promising and were reported in the last report. However, as we have learned more about the L1 assay, we believe that those preliminary results were an artifact caused by the stimulatory influence of the mutant p53 causing the cells to grow faster. To some extent this is also a function of cell plating

density and whether the G418 selection for neomycin resistance is able to be effective before the cells approach confluence. Ultimately, after many repetitions, we can see no influence of p53 mutation on the L1 retrotransposition rate. However, we also wanted to look at the effect of cell cycle in general and we have been able to demonstrate that slowing cell growth by a factor of two by lowering the growth temperature results in an order of magnitude decrease in retrotransposition rates. Furthermore, this effect correlates with growth rate and not just temperature. If the temperature is lowered just at the beginning of the assay, the rate does not change. Thus, the L1 enzymes are not susceptible to temperature, instead, lowering the temperature for a prolonged period has a secondary effect that greatly lowers retrotransposition rates. We have utilized fluctuation analysis on long-term transformants for all of these assays and have also created a transient transfection-based assay. At this point we are gearing up to look at various breast cancer cell lines for their retrotransposition potential, as well as cells with various genetic defects associated with tumorigenesis and DNA repair. Thus, although we cannot yet answer the question of whether transformation alters retrotransposition and therefore retrotransposition may contribute to the progression in cancer, we now have the tools and should be able to test a number of model systems soon.

(7) Key Research Accomplishments

Year 1

- Establishment of optimum conditions for amplification of the most recent subfamilies of Alu inserts
- Obtaining clear displays of the Ya8 subfamily on acrylamide and agarose gels which allow the isolation of insertion polymorphisms between different individuals.
- Demonstrating the use of modified primers that display subsets of the Ya5 elements that will allow at least a substantial portion of Ya5 inserts to be studied.

Year 2

- Identification of the youngest, most active Alu subfamilies that can be amplified and displayed directly without the use of subtraction protocols.

Year 3

- Development of a complete understanding of the recent amplification of Alu elements in the human genome based on the fusion of bioinformatics on the complete human genome sequence and laboratory-based studies.
- Development of approaches to use retroposition reporter gene systems for studies of the role of various genes and environmental influences on the retrotransposition frequency.

(8) Reportable Outcomes

Astrid Roy-Engel was supported by this grant.

- **Deininger, P.** and Batzer, M. (1999) *Alu repeats and human disease*. Mol Gen and Metab **67**, 183-193.
- Roy, A.M., M. Carroll, D.H. Kass, Sun, MA. Batzer, **P.L. Deininger** (1999) *Recently integrated human Alu repeats: Finding needles in the haystack*. Genetica **107**, 149-61.
- Roy, A.M., M.L. Carroll, S.V. Nguyen, A.-H. Salem, M. Oldridge, A.O.M. Wilkie, M.A. Batzer, and **P. L. Deininger** (2000) *Potential gene conversion and source gene(s) for recently integrated Alu elements*. Genome Research **10**, 1485-1495.
- Roy-Engel, ML Carroll, E. Vogel, RK Garber, SV Nguyen, A-H Salem, MA Batzer and **P. Deininger** (2001) *Alu insertion polymorphisms for the study of human genomic diversity*. Genetics (in press)
- ML. Carroll, A. Roy-Engel, SV. Nguyen, A-H Salem, E. Vogel, B. Vincent, J. Myers, Z. Ahmed, L. Nguyen, M. Sammarco, WS. Watkins, J. Henke, W. Makalowski, LB. Jorde, **P. Deininger**, and MA. Batzer. (2001) *Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity*. J. Mol. Biol. (in press).

(9) Conclusions

We were able to develop a PCR procedure that can selectively amplify the subset of most recently inserted Alu elements. Although we were able to display a subset of these elements, we were unable to overcome sufficient technical difficulties to allow an assessment of the number of Alu insertions occurring in breast tumors.

We developed quantitative approaches to measure the retrotransposition capability of different cell types using a reporter-gene approach. Using this approach we showed that dominant negative p53 mutations did not alter retrotransposition rates, but that major changes to cells influencing growth rates had a tremendous influence. We are currently gearing up for a full assessment of breast cancer cell lines, and a number of genes associated with tumorigenesis using this quantitative assay.

Reference List

1. Deininger, P L. and Batzer, M A. Alu repeats and human disease. *Mol Genet Metab* 67, 183-193. 1999.
Ref Type: Abstract
2. M. Batzer et al., *Nucleic Acids Res.* 19, 3619-3623 (1991).
3. P. Deininger and V. Slagel, *Mol.Cell.Biol.* 8, 4566-4569 (1988).
4. J. V. Moran et al., *Cell* 87, 917-927 (1996).

APPENDIX

one reprint for:

- Deininger, P. and Batzer, M. (1999) *Alu repeats and human disease*. *Mol Gen and Metab* **67**, 183-193.
- Roy, A.M., M. Carroll, D.H. Kass, Sun, MA. Batzer, P.L. Deininger (1999) *Recently integrated human Alu repeats: Finding needles in the haystack*. *Genetica* **107**, 149-61.
- Roy, A.M., M.L. Carroll, S.V. Nguyen, A.-H. Salem, M. Oldridge, A.O.M. Wilkie, M.A. Batzer, and P. L. Deininger (2000) *Potential gene conversion and source gene(s) for recently integrated Alu elements*. *Genome Research* **10**, 1485-1495.
- Roy-Engel, ML Carroll, E. Vogel, RK Garber, SV Nguyen, A-H Salem, MA Batzer and P. Deininger (2001) *Alu insertion polymorphisms for the study of human genomic diversity*. *Genetics* (in press)
- ML. Carroll, A. Roy-Engel, SV. Nguyen, A-H Salem, E. Vogel, B. Vincent, J. Myers, Z. Ahmed, L. Nguyen, M. Sammarco, WS. Watkins, J. Henke, W. Makalowski, LB. Jorde, P. Deininger, and MA. Batzer. (2001) *Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity*. *J. Mol. Biol.* (in press).

Potential Gene Conversion and Source Genes for Recently Integrated Alu Elements

Astrid M. Roy,^{1,6} Marion L. Carroll,^{2,6} Son V. Nguyen,² Abdel-Halim Salem,² Michael Oldridge,³ Andrew O. M. Wilkie,^{3,4} Mark A. Batzer,^{2,7} and Prescott L. Deininger^{1,5,7,8}

¹Tulane Cancer Center, Department of Environmental Health Sciences, Tulane University Medical Center, New Orleans, Louisiana 70112, USA; ²Departments of Pathology, Biometry and Genetics, Biochemistry, and Molecular Biology, Stanley S. Scott Cancer Center, Neuroscience Center of Excellence, Louisiana State University Health Sciences Center, New Orleans, Louisiana 70112, USA; ³Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX2 6HE, UK; ⁴Oxford Craniofacial Unit, The Radcliffe Infirmary NHS Trust, Oxford OX2 6HE, UK; ⁵Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, New Orleans, Louisiana 70121, USA

Alu elements comprise >10% of the human genome. We have used a computational biology approach to analyze the human genomic DNA sequence databases to determine the impact of gene conversion on the sequence diversity of recently integrated Alu elements and to identify Alu elements that were potentially retroposition competent. We analyzed 269 Alu Ya5 elements and identified 23 members of a new Alu subfamily termed Ya5a2 with an estimated copy number of 35 members, including the de novo Alu insertion in the *NFI* gene. Our analysis of Alu elements containing one to four (Yal-Ya4) of the Ya5 subfamily-specific mutations suggests that gene conversion contributed as much as 10%–20% of the variation between recently integrated Alu elements. In addition, analysis of the middle A-rich region of the different Alu Ya5 members indicates a tendency toward expansion of this region and subsequent generation of simple sequence repeats. Mining the databases for putative retroposition-competent elements that share 100% nucleotide identity to the previously reported de novo Alu insertions linked to human diseases resulted in the retrieval of 13 exact matches to the *NFI* Alu repeat, three to the Alu element in *BRCA2*, and one to the Alu element in *FGFR2* (Apert syndrome). Transient transfections of the potential source gene for the Apert's Alu with its endogenous flanking genomic sequences demonstrated the transcriptional and presumptive transpositional competency of the element.

Alu elements belong to a class of retroposons termed SINEs. SINEs are Short INterspersed Elements usually ~100–300 bp in length commonly found in introns, 3' untranslated regions of genes, and intergenic genomic regions (Deininger and Batzer 1993). Alu is the most abundant class of SINEs in primate genomes, reaching a copy number in excess of one million/haploid genome (Jelinek and Schmid 1982; Jurka et al. 1993; Smit 1999). Alu elements increase their genomic copy number by an amplification process termed retroposition (Rogers and Willison 1983; Weiner et al. 1986).

Alu elements appear to have arisen in the last 65 million years (Deininger and Daniels 1986). The human Alu family of repeats is composed of a small number of distinct subfamilies characterized by subfamily-specific diagnostic mutations (Slagel et al. 1987; Willard et al. 1987; Shen et al. 1991; Batzer et al. 1996b). The source Alu gene(s) for each of the subfami-

lies has been retropositionally active during different periods of primate evolution. The rate of Alu amplification (mostly Sx subfamily) appears to have reached its peak between 60 and 35 million years, and subsequently decreased several orders of magnitude to the present amplification rate (Shen et al. 1991). Only a limited number of SINEs, termed master or source genes, appear to be capable of retroposition (Deininger and Daniels 1986; Batzer et al. 1990; Deininger et al. 1992), although the critical factor(s) defining functional source genes are not understood. A variety of factors influence the retroposition process (Schmid and Maraia 1992). All of the recently integrated young Alu subfamilies appear to be retropositionally active. Almost all of the recently integrated Alu elements within the human genome belong to one of four closely related subfamilies (Y, Ya5, Ya8, and Yb8), with the majority being Ya5 and Yb8 subfamily members (Batzer et al. 1990, 1995; Deininger and Batzer, 1999).

Previously, analysis of individual Alu elements from the different subfamilies involved laborious procedures, such as cloning, library screening, and subsequent sequencing (Batzer et al. 1990, 1995; Arcot et al. 1995a). However, the availability of large-scale human

⁶These authors contributed equally to this work.

⁷These authors contributed equally to this work as senior authors.

⁸Corresponding author.

E-MAIL PDEININ@TCS.TULANE.EDU ; FAX (504) 588-5516.

Article and publication are at www.genome.org/cgi/doi/10.1101/gr.152300.

genomic DNA sequences as a result of the Human Genome Project facilitates genomic database mining for Alu elements (Roy et al. 1999). We have taken advantage of these databases and have analyzed a significant portion of the Alu Ya5 subfamily, as well as intermediates between the Ya5 subfamily and the ancestral Alu Y subfamily. In addition, we searched the databases for putative retroposition-competent source Alu genes that generated the de novo Alu inserts associated with a number of human diseases (Deininger and Batzer 1999).

RESULTS

Computational Analyses

To search for subfamilies unidentified previously within the Ya5 Alu subfamily, we selected all of the Alu family members that matched our Ya5 consensus query sequence from the human genome non-redundant (nr) database. Only Ya5 elements found randomly within other sequences were included in our analysis, thereby eliminating Alu elements that had been identified previously in directed Alu-specific projects. In addition, truncated Alu elements were

eliminated from the analysis. Ya4 elements that did not contain the first Ya5-specific diagnostic mutation #11 (Fig. 1) (Shen et al. 1991), which is a CpG dinucleotide in the Ya5 subfamily, were considered as Ya5 Alu family members. We obtained a total of 269 matches to the Ya5 query sequence that met our criteria. Of these, 47 shared 100% nucleotide identity with the subfamily consensus sequence and 83 were near perfect matches (aside from a few CpG mutations).

Analysis of the 269 Ya5 Alu elements resulted in the initial identification of two subsets of potential subfamilies containing two diagnostic mutations each, one with six members and the other with four. These subfamilies will be referred to as Ya5a2 and Ya5b2, respectively, in compliance with the standard Alu subfamily nomenclature (Batzer et al. 1996a). Each consensus sequence with the two diagnostic mutations specific to each new Alu subfamily is shown in Figure 1. Interestingly, the de novo Alu Ya5 insert present within an intron of the *NF1* gene (Wallace et al. 1991) is an exact match to the Ya5a2 consensus. The nr database contained 16.0% of human DNA sequences for a total of 515,596,000 bases on the date of the search. The estimated size of the Ya5a2 subfamily is $(3 \times 10^9$ bp/515,596,000 bp) \times 6 unique Ya5a2 matches = 35 subfamily members. In comparison, the estimated size of the Ya5b2 subfamily is $(3 \times 10^9$ bp/515,596,000 bp) \times 4 unique Ya5b2 matches = 22 subfamily members. We utilized only the randomly found Ya5a2 elements for the calculations to avoid overestimating the size of the subfamilies. However, these numbers may be underestimations, because some specific polymorphic elements of these subfamilies may not be represented in the database.

bp/515,596,000 bp) \times 6 unique Ya5a2 matches = 35 subfamily members. In comparison, the estimated size of the Ya5b2 subfamily is $(3 \times 10^9$ bp/515,596,000 bp) \times 4 unique Ya5b2 matches = 22 subfamily members. We utilized only the randomly found Ya5a2 elements for the calculations to avoid overestimating the size of the subfamilies. However, these numbers may be underestimations, because some specific polymorphic elements of these subfamilies may not be represented in the database.

To derive a second estimate of the copy numbers of the Ya5a2 and Ya5b2 Alu subfamilies, we used their consensus sequences as queries for the high throughput genome sequence (htgs) and genomic survey sequence (gss) databases. Seventeen additional Alu Ya5a2 elements were found in these searches. Of the 23 total Ya5a2 elements, 13 shared 100% nucleotide identity with the subfamily consensus sequence. No additional Ya5b2 elements were found in the other databases, therefore the Ya5b2 subfamily was not subjected to further analysis. Three additional potential subfamilies, Ya5a1 (five members), Ya5b1 (four members), and Ya5c1 (four members) with only one specific diagnostic mutation were identified (Fig. 1). Because of the small copy number, and the possibility that some

Ya5	GGCCGGCGCGGTGGCTCACGCCCTGAAATCCCAGCAGCTTGGGAGGCCGAGGCCGGCGGA	60
Ya5a2	60
Ya5b2	60
Ya5a1	60
Ya5b1	60
Ya5c1	60
Ya5	TCACGAGGTCAAGGAGATCGAGACCATCCGGCTAAACGGTAAACCCCGTCTACTAA	120
Ya5a2A.....	120
Ya5b2A.....C.....	120
Ya5a1G.....	120
Ya5b1	120
Ya5c1	120
Ya5	AAATACAAAAAA-TTAGCCGGGCTAGTGGCGGGCCGTAGTCCCAGCTACTGGGAG	179
Ya5a2A.....	180
Ya5b2	179
Ya5a1	179
Ya5b1G.....	179
Ya5c1	179
Ya5	13.....14.....15.....	
Ya5a2A.....	180
Ya5b2	179
Ya5a1	179
Ya5b1G.....	179
Ya5c1	179
Ya5	GCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCCGAGCTTGCAGTGAGCCGAGATCCCG	239
Ya5a2	240
Ya5b2	239
Ya5a1	239
Ya5b1G.....	239
Ya5c1	239
Ya5	CCACTGCACTCCAGCCTGGCGACAGAGCGAGACTCCGTCTC	281
Ya5a2	282
Ya5b2	281
Ya5a1	281
Ya5b1	281
Ya5c1	281

Figure 1 Consensus sequence alignment of Ya5, and the potential new subfamily members identified. Nucleotide substitutions at each position are indicated with the appropriate nucleotide. Deletions are marked by dashes (-). The Ya5 diagnostic nucleotides are indicated in bold with the corresponding diagnostic number above as defined by Shen et al. (1991).

of those represent parallel mutations rather than new subfamilies, no further analyses were performed.

To determine the age of the Ya5a2 subfamily, we divided the nucleotide substitutions within the elements into those that have occurred in CpG dinucleotides and those that have occurred in non-CpG positions. The distinction between types of mutations is made because the CpG dinucleotides mutate at a rate that is ~10 times faster than non-CpG (Labuda and Striker 1989; Batzer et al. 1990), as a result of the deamination of 5-methylcytosine (Bird 1980). A total of five non-CpG mutations and seven CpG mutations occurred within the 23 Alu Ya5a2 subfamily members identified. By use of a neutral rate of evolution for primate-intervening DNA sequences of 0.15%/one-million years (Miyamoto et al. 1987) and the non-CpG mutation rate of 0.092% (5/5382 bases using only non-CpG bases) within the 23 Ya5a2 Alu elements, yields an estimated average age of 0.62 million years for the Ya5a2 subfamily members with a predicted 95% confidence level in the range of 0.28–1.08 million years, given that the mutations were random and fit a binomial distribution. The Ya5a2 subfamily appears to be much younger than Ya5, Ya8, or Yb8 Alu subfamilies with estimated ages of 2.8 million years (Batzer et al. 1990), 2.75 million years (Roy et al. 1999), and 2.7 million years (Batzer et al. 1995), respectively (Fig. 2).

Determination of the number of elements that perfectly match the subfamily consensus sequence can also give an indirect estimate of Alu subfamily age and recent rate of mobilization. Recently transposed Alu

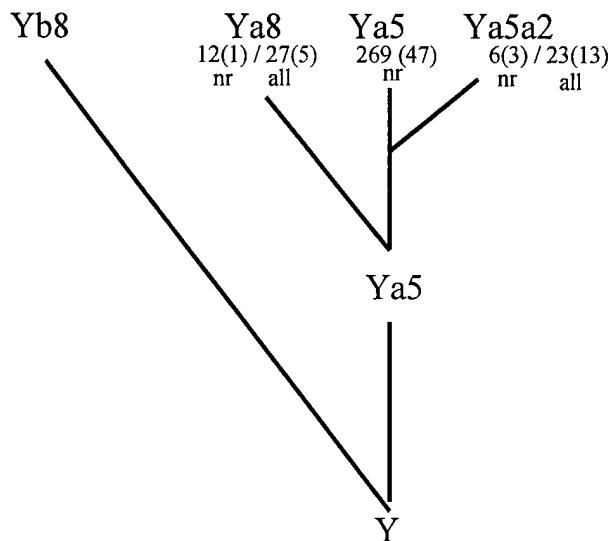


Figure 2 Schematic for the evolution of recently integrated Alu subfamilies. The origin of the Ya5a2 Alu subfamily is shown after the divergence of Ya5 and Yb8 elements. The total number of elements found in the nr-database (perfect matches in parenthesis) are shown first separated by a slash from the total number of elements found in all three databases (nr, gss, htgs). For the Ya5 elements only the nr-database results are shown.

Table 1. Alu Middle A-Rich Region

Ya5-middle A-rich region	A _n							
	4	5	6	7	8	9	10	11
T(A _n)TACA ₆ TT ^a	0	269 ^c	9	1	0	1	—	—
TA ₅ TAC(A _n)TT ^b	0	2	269 ^c	37 ^d	11	7	3	0

^an = 5 in Ya5 consensus.

^bn = 6 in Ya5 consensus.

^cData from the non-redundant database only.

^dAll 23 Ya5a2 members are included.

elements share higher levels of nucleotide identity with their source copies because they have not resided in the genome long enough to accumulate random mutations. In contrast, older Alu elements that have resided in the genome for longer periods of time tend to have less nucleotide identity with their source genes as a result of the accumulation of random mutations subsequent to integration into the genome. We compared our search results for the Ya5a2 subfamily with parallel searches from the Ya8 and Ya5 Alu subfamilies. Our BLAST searches from the nr database yielded one perfect match of 12 elements for Ya8, 47 of 269 for Ya5, and 3 of 6 for Ya5a2 (Fig. 2). Searching all three databases (nr, gss, and htgs) yielded 5 perfect matches of 27 for Ya8 and 13 of 23 for Ya5a2. These results are in good agreement with the previous estimates, indicating that Ya5a2 is the youngest Alu subfamily reported to date, as it also has the highest proportion of elements that share 100% nucleotide identity with the consensus sequence.

Stability of the Middle A-Rich Region in Alu Ya5 Members

The oligo-dA-rich tails and middle A-rich regions of Alu elements have been shown previously to serve as nuclei for the genesis of simple sequence repeats (Arcot et al. 1995b). In the autosomal recessive neurodegenerative disease, Friedreich ataxia, the most common mutation, is the hyperexpansion of a GAA within the middle A-rich region of an Sx Alu element (Montermini et al. 1997). Because these regions appear unstable, we analyzed the middle A-rich region of Alu elements retrieved from the databases to detect expansions/contractions of this sequence.

To evaluate potential expansions/contractions, we performed a BLAST query of three databases (nr, htgs, and gss) using the Alu Ya5 consensus sequence with varying numbers of A nucleotides within the middle A-rich region (TA_nTACA₆TT). Our results demonstrate that the majority of the elements identified matched the consensus sequence. However, there is a trend for an A expansion at both positions (Table 1). In contrast,

very few sequence contractions were detected for any of the positions.

Human Genomic Variation

To determine the human genomic variation associated with the Ya5a2 Alu subfamily members, we selected the 13 Ya5a2 elements identical to the subfamily consensus sequence as well as 2 others and determined the degree of fixation associated with the elements using PCR-based assays of a panel of diverse human DNA samples with the primers shown in Table 2. The panel is composed of 20 individuals of European origin, African-Americans, Greenland natives, and Egyptians for a total of 80 individuals (160 chromosomes). The Alu elements were classified as fixed absent, fixed present, and high, intermediate, or low frequency insertion polymorphisms (see Table 3 for definitions). By use of this approach, 3 of the 14 elements tested (Ya5NBC206, Ya5NBC207, and Ya5NBC235) were always present in the human genomes that were surveyed, suggesting that these elements became fixed in the genome prior to the radiation of modern humans from Africa. Five of the elements (Ya5NBC208, Ya5NBC240, Ya5NBC241, Ya5NBC242, and Ya5NBC220) are intermediate frequency Alu insertion polymorphisms. The remaining six elements are low-frequency Alu insertion polymorphisms (Table 3). The population-specific genotypes and levels of heterozygosity for each element are shown in Table 4. The high proportion of polymorphic elements is in good agreement with our other observations, indicating that

the Ya5a2 subfamily is younger than any of the other Alu subfamilies identified previously in the human genome.

Gene Conversion and Alu Sequence Diversity

In our query of the human genome (nr) database, 91 of the Alu elements identified contain one to four of the five Ya5 diagnostic nucleotides (Fig. 1). Of these 91 intermediate elements, 4 are Ya1, 1 Ya2, 7 Ya3, and 79 Ya4 Alu elements (Fig. 3). Surprisingly, not all of the Alu elements with different numbers of subfamily mutations had the same combination of mutations. To facilitate identification of the individual elements with different diagnostic mutation combinations, the diagnostic nucleotides were numbered consecutively in order of abundance (Ya3.1, Ya3.2, etc., see Fig. 3). Seventeen Alu elements (Ya4.4) did not contain the first diagnostic mutation (#11), but were still classified as Ya5 for the analyses outlined above.

Previous evolutionary analyses of the Ya5 founder element with different primate DNA samples demonstrated the sequential accumulation of the Ya5 diagnostic mutations with diagnostic positions #13/#14 first, followed by #12/#16, and finally position #11 (Shaikh and Deininger 1996). Our data are not consistent with a sequential order in the accumulation of the diagnostic mutations. The elements classified as Ya1, Ya2, Ya3.4, Ya3.5, and Ya4.4 (26 total) fit the proposed order (Fig. 3). However, the remaining 65 elements represent almost every other permuted order. Several mechanisms could explain the occurrence for mosaic

Table 2. Alu Ya5a2 PCR Primers, Chromosomal Locations, and PCR Product Sizes

Name	5' Primer sequence (5'-3')	3' Primer sequence (5'A-3')	A.T. ^a	Chromo- some ^b	Product size ^c	
					filled	empty
Ya5NBC206	TCCTTAGCTATCTACAAGCTACAT	ACACATTTCTCAAGAGGTCAAAG	60°C	4	734	424
Ya5NBC207	CAGTTTATACACTGGCTGTTTC	TTGTTAGGAGAAAGAGGGAAATCT	50°C	6	443	122
Ya5NBC208	AATACCTGTACATCTCACCCCTA	TCTCTCTGCTGCACAGTTGTT	50°C	14	441	115
Ya5NBC240	CAGGAGATAAAATGTTGGAGAGT	TAACCTGGACAGTGAGTTTACCTG	55°C	9	505	202
Ya5NBC241	GGTTCCAATAGAGAGCAACAGAA	ACCTTAAGCTTCCCCAGA	55°C	15	392	66
Ya5NBC242	AACAAAATCCCTTCTCCA	GGCAATCTGACCTTGGTAA	55°C	7	503	192
Ya5NBC7	TGATGGATATGGGTTGGTTC	GGACTGAAACTAGTTCAACCATTGTC	60°C	7	522	216
Ya5NBC205	ACATGAAGGGCGACTGTAT	TGCTGCTGCATTATCAACTG	50°C	21	435	81
Ya5NBC209	GTCTATGGGAAGATGAAGAATAGGA	GATGGAGTCACTCATGTGAAAGTA	55°C	14	447	116
Ya5NBC239	CAGCTGAGAACTGTACAATAGAA	ATCAATGACTGACTTGTGCTGAGT	55°C	9	531	198
Ya5NBC243	CCATGATTCTGTATTACCA	AGGAGACCTGCAATGAATG	60°C	21	406	86
Ya5NBC220	AAATCAAGCTGCCATACCTCA	GAAACCATCTTACACTGCG	60°C	1	463	141
Ya5NBC235	CCCAAGGCACTTGCTGTTA	CCCTTCGAGAAAGAGGAAGG	50°C	2	391	76
Ya5NBC244	CCTATGGCTGAAACTCTGAAACT	ATATCTGGCTCACTAGACAAGCAC	60°C	18	453	130
Ya5NBC237 ^d	CCCATGGAGGGTCTTCCTA	CTGAAACCACCTTCACAGT	60°C	1	410	88

^aAmplification of each locus required 2.5 min at 94°C initial denaturing, and 32 cycles for 1 min 94°C, 1-min annealing temperature (A.T.) and 1-min elongation at 72°C. A final extension time of 10 min at 72°C was also used.

^bChromosomal location determined from accession information or by PCR analysis of NIGMS monochromosomal hybrid cell line DNA samples.

^cEmpty product sizes calculated by removing the Alu element and one direct repeat from the filled sites that were identified.

^dAlu Ya5a2 element of the *FGFR2* gene.

Table 3. Alu Ya5a2 (*NF1*)-Associated Human Genomic Diversity

Ya5a2 elements	Accession no. (duplicates)	Position	Allele frequency ^a
Ya5NBC206	AC004057	76767-77048	fixed present
Ya5NBC207	AL118555 (AL132992)	9981-9700 (40728-41009)	fixed present
Ya5NBC208	AL109919	70170-69889	intermediate
Ya5NBC220	AC007611	136715-136434	intermediate
Ya5NBC240	AC133410 (AL135841)	34800-35081 (49829-49548)	intermediate
Ya5NBC241	AC018924	144017-144298	intermediate
Ya5NBC242	AC009517	161301-161582	intermediate
Ya5NBC7	AC004848	24522-24241	low
Ya5NBC205	AL011328	204488-204207	low
Ya5NBC209	AC00808	147056-146775	low
Ya5NBC239	AL133284	115867-115586	low
Ya5NBC244	AC026839	64885-64604	low
Ya5NBC243	AJ011929	151192-151473	low
Ya5NBC235 ^b	AQ748733	458-739	fixed present
Ya5NBC237 ^c	AL031274	33175-33501	intermediate

^aAllele frequency was classified as fixed present, fixed absent, low, intermediate, or high frequency insertion polymorphism. (Fixed present) every individual tested had the Alu element in both chromosomes; (low frequency insertion polymorphism) the absence of the element from all individuals tested, except for one or two homozygous or heterozygous individuals; (intermediate frequency insertion polymorphism) the Alu element is variable as to its presence or absence in at least one population; (high frequency insertion polymorphism) the element is present in all individuals in the populations tested, except for one or two heterozygous or absent individuals.

^bSeveral Ns.

^cYa5NBC237 is the exact match to the *FGFR2* Alu insertion.

Alu elements, which are addressed in the discussion section. However, we believe the most likely explanation for the existence of these mosaic elements is through gene conversion events. A limited amount of gene conversion between Yb8 Alu elements has been

reported previously (Batzer et al. 1995; Kass et al. 1995). In theory, gene conversion may change the sequence of all or part of any Alu element in either an evolutionarily forward (Ya5 subfamily in this case) or backward (Y subfamily) direction by changing the di-

Table 4. Alu Ya5a2-Associated Human Genomic Diversity

Elements	African American		Greenland natives		European		Egyptian		
	genotype ^a	fAlu ^b	genotypes	fAlu	genotypes	fAlu	genotypes	fAlu	het. ^c
Ya5NBC206	20 0 0	1.000	20 0 0	1.000	20 0 0	1.000	20 0 0	1.000	0.000
Ya5NBC207	20 0 0	1.000	20 0 0	1.000	20 0 0	1.000	20 0 0	1.000	0.000
Ya5NBC208	4 1 7	0.375	3 0 4	0.429	13 0 6	0.684	7 0 5	0.583	0.482
Ya5NBC236	5 6 2	0.615	5 8 6	0.474	15 5 0	0.875	6 8 1	0.667	0.422
Ya5NBC240	5 1 9	0.367	11 0 4	0.733	5 1 10	0.344	5 3 3	0.591	0.464
Ya5NBC241	3 9 5	0.441	6 11 2	0.605	0 7 11	0.194	3 8 4	0.467	0.459
Ya5NBC242	2 13 1	0.531	7 4 3	0.643	3 4 11	0.278	3 3 1	0.643	0.474
Ya5NBC7	0 0 19	0.000	0 0 20	0.000	0 0 20	0.000	0 0 20	0.000	0.000
Ya5NBC205	0 0 20	0.000	0 0 20	0.000	0 0 20	0.000	0 0 20	0.000	0.000
Ya5NBC209	0 1 17	0.028	0 0 17	0.000	0 0 19	0.000	0 0 19	0.000	0.000
Ya5NBC239	0 0 20	0.000	0 0 20	0.000	0 0 20	0.000	0 0 20	0.000	0.000
Ya5NBC243	0 0 20	0.000	0 0 20	0.000	0 0 20	0.000	0 0 20	0.000	0.000
Ya5NBC220	0 14 5	0.368	1 15 2	0.472	0 18 1	0.474	0 9 2	0.409	0.502
Ya5NBC244	0 0 12	1.000	— — —	—	0 0 10	0.000	0 0 8	0.000	0.000
Ya5NBC235	20 0 0	1.000	20 0 0	1.000	20 0 0	1.000	20 0 0	1.000	0.000
Ya5NBC237 ^d	18 1 0	0.974	15 4 0	0.895	20 0 0	1.000	18 1 0	0.974	0.075

^aGenotypes: +/+ Alu, +/− Alu, −/− Alu.

^bFrequency of the presence of the Alu.

^cAverage heterozygosity.

^dYa5NBC237 is the exact match to the *FGFR2* Alu insertion.

— not determined.

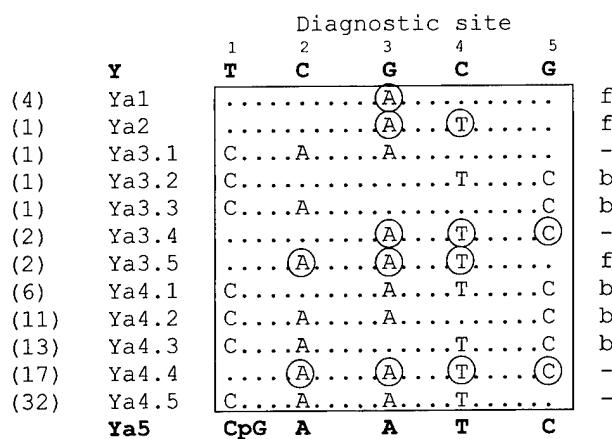


Figure 3 Evolution of the diagnostic nucleotide positions from Y to Ya5 Alu elements. Alignment of the five Alu Ya5 diagnostic nucleotides as defined by Shen et al. (1991) and the different Ya1, Ya2, Ya3, and Ya4 elements found in the nr database. For easy reference, individual elements containing different combinations of the diagnostic mutations were numbered consecutively in order of abundance (Ya3.1, Ya3.2, etc.). Ya4.4 elements were considered as Ya5 elements in the first Ya5 subfamily analysis in this paper. The total number of elements found for each subgroup is indicated at left in parenthesis. Potential forward (f) or backward (b) gene conversions are indicated at right. The previously reported order of appearance of Ya5 diagnostic mutations (Shaikh and Deininger 1996) is indicated below. Elements with diagnostic mutations that follow the stepwise hierarchical accumulation are circled.

agnostic mutations. In addition, double gene conversions would be extremely rare, making the direction of the gene conversion clear in some elements. We classified the 91 mosaic Alu element sequences as gene converted forward (f), backward (b), or could not be determined (-). (see Fig. 3) If the Alu elements that fit the proposed sequential evolution are ignored in the analysis, all of the other elements may be classified as backward gene conversion (32 total) or could not be determined (33 total), and none were clearly gene-converted forward. Therefore, backward gene conversion may have contributed to between 10% and 20% (32 to 65/269 Ya5 + [91-17] Ya1-Ya4) of the Alu Ya5 sequence diversity. Interestingly, evaluation of the five random Ya5a2 non-CpG mutations shows that one mutation in position #13 is a backward mutation to the Y subfamily, another putative example of a reverse gene conversion.

In Search of Retroposition-Competent Alu Repeats

Sixteen different Alu insertions have been linked to human diseases (Deininger and Batzer 1999). Four belong to the Alu Y subfamily, one to the Ya4 subfamily, eight to the Ya5 subfamily, and three to the Yb8 subfamily. Closer inspection of the nucleotide sequences of these Alu elements show that they have some mutations that are different from their respective subfamily consensus sequences. Because these Alu insertions

are very recent in origin, they are likely to be identical to their source genes aside from rare mutations introduced during reverse transcription of the Alu element. Therefore, sequence database queries utilizing each Alu element along with its individual mutations (away from the subfamily consensus sequence) may facilitate the identification of the source Alu element that generated the copy. This strategy is similar to that used previously in the identification of active LINE elements from the human genome (Dombroski et al. 1993).

A database query using the sequence of the individual Alu elements responsible for each disease to mine three databases (nr, htgs, and gss) identified exact complements to four of the disease-associated Alu repeats. Thirteen of the identified elements were exact matches to the *NF1* Alu insertion (Ya5a2 subfamily, Table 3; Wallace et al. 1991); three were exact matches to the *BRCA2* Alu element (Miki et al. 1996) (accession nos. AL121964, AL136319, and AL135778); one matched the *FGFR2* Alu repeat (Oldridge et al. 1999) (accession no. AL031274); and one matched the Alu repeat in the *IL2RG* gene (Lester et al. 1997) (accession no. AC010888).

Potential Source Gene for the Ya5 Insert in *FGFR2*

As mentioned above, our BLAST query only detected one exact match (accession no. AL031274 or Ya5NBC237) to the Ya5 Alu found in the *FGFR2* gene that caused Apert syndrome. We estimated the level of human genomic variation associated with Ya5NBC237 using the same human DNA panel and determined that it was an intermediate frequency Alu insertion polymorphism (Table 4).

Mobilization-competent Alu elements must be capable of transcription, the first step in the retroposition process. To evaluate Alu Ya5NBC237 as a potential source gene for the de novo insert in the patient with Apert syndrome, we determined its transcription capability. Constructs with the genetic loci containing the Ya5NBC237 Alu and the de novo Apert syndrome Alu element were made. Transcription levels from the two constructs were evaluated by Northern blot analysis relative to a control plasmid in which the Alu element is flanked immediately upstream by vector sequence.

Transient transfections (Fig. 4) of the constructs into rodent cell line C6 (rat glial tumor) were performed. Although the Alu element in the control plasmid has an intact internal Pol III promoter, Alu transcripts are barely detectable from the control plasmid. In contrast, the transcription from the Apert's Alu element and its potential source gene were elevated three- to fourfold, as expected for putative mobilization-competent Alu repeats. This result suggests that the genomic flanking sequence of Ya5NBC237 probably makes the Alu transcription competent, one of the several requirements of a source gene. The same results

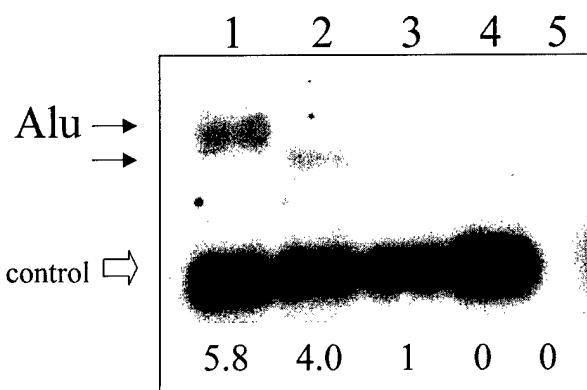


Figure 4 Evaluation of transcriptional capability of the potential *FGFR2* source Ya5 Alu element. The transcriptional efficiency of the de novo *FGFR2* Alu repeat and its putative source gene were evaluated by Northern blot analysis from transient transfection studies. The following constructs were evaluated: (lane 1) p²⁹⁰Ap, (lane 2) p⁻⁴¹⁶Ya5NBC237, and (lane 3) p^{NP}Ya5NBC237. Lanes 4 and 5 are internal control only, and no DNA controls, respectively. Small arrows indicate the Alu transcripts and the open arrow indicates the internal control transcript. The ratio of the Alu transcript/control transcript (numbers below) was normalized to the p^{NP}Ya5NBC237 transcription ratio, which was assigned the arbitrary value of 1.

were obtained from transfections in the human embryonic kidney cell line 293 (data not shown).

DISCUSSION

Our computational and experimental analyses of the Ya5 subfamily of Alu repeats provides an overall picture of the most active of the recently integrated young Alu subfamilies from the human genome. The analysis of Alu Ya5 repeats allowed us to address a number of questions about the biology of these elements, such as the potential impact of gene conversion events, and the identification of Alu family members from the human genome that may be capable of retroposition.

Alu elements spread throughout the genome by retroposition in the last 65 million years. The master/source gene model (Batzer et al. 1990; Shen et al. 1991; Deininger et al. 1992) posits that a very small subset of the >1,000,000 Alu elements within the human genome are capable of high levels of retroposition; although a much larger number may make a few copies. The formation of Alu subfamilies may be explained by the sequential accumulation of mutations within the active source gene(s) followed by proliferation of the mutated source elements. A number of studies indicate that relatively few source Alu genes have played a dominant role in the amplification and evolution of Alu elements (Shen et al. 1991; Deininger et al. 1992; Deininger and Batzer 1993; Kapitonov and Jurka 1996). Although retroposition is the primary mode of SINE mobilization and sequence evolution through

mutations in the source gene(s), our analysis suggests that gene conversion and genetic instability of Alu-based simple sequence repeats have also had a significant impact on the sequence architecture of this major family of human genomic sequences.

There are several alternatives that could explain the occurrence of mosaic Alu elements. First, some of the mosaic Alu elements with a single mutation could be explained by the occurrence of parallel mutations. However, this seems unlikely unless there were selection for these specific mutations, possibly through a post-transcriptional selection process (Sinnott et al. 1992). It is also difficult to envision a selection process that would only select for mutations at adjacent diagnostic positions, such as we see here. Also, recombination between different Alu elements could have generated some of these intermediate Alu elements that contain a mosaic of diagnostic mutations. However, in many cases, multiple recombination events would be required to obtain this outcome, making it highly unlikely. Although there are alternative mechanisms, we believe gene conversion is the most likely explanation for the occurrence of mosaic Alu elements.

The mechanisms of genome-wide gene conversion between mobile elements are not well understood in humans (see Kass et al. 1995, and references therein). Our data show that even the very short, dispersed Alu elements appear to be capable of high levels of gene conversion, which usually involve only short sequence stretches. In addition, our data show that reverse or backward gene conversions may be more favored. It seems likely that higher levels of the Y element copy number (Shen et al. 1991) or transcription (Shaikh et al. 1997) may play a role in determining the directionality of the gene conversion events. Although older Alu subfamilies, such as J and Sx are present in higher copy numbers in the genome, they diverged greatly from their consensus sequences due to mutations that have accumulated throughout evolution. Gene conversion would not be favored between such divergent sequences. However, Alu Y elements tend to be more conserved (better matches to Ya5) and with high copy number (Batzer et al. 1995). Therefore, both abundance (genomic copy number and/or transcript levels) and sequence identity appear to be influential in the Alu gene conversion events observed.

There are multiple examples of gene conversion events in literature. Genetic exchange between exogenous and different endogenous mouse L1 elements has been demonstrated previously to readily occur (Belmaaza et al. 1990). Kass et al. (1995) reported previously a gene conversion event in which one of the oldest Alu family members was converted to one of the youngest Alu subfamilies, Yb8. In addition, a partially converted Yb8 Alu element was also reported previously by Batzer et al. (1995). In yeast, some types of

mobile elements spread through the genome by gene converting pre-existing elements (Hoff et al. 1998). When we combine this type of mobilization in the yeast genome with the Alu gene conversions reported previously, as well as those in this paper, one could argue that gene conversion may represent a second type of amplification mechanism for short interspersed elements in the human genome. These observations suggest that evolutionary studies of all types of interspersed elements that ignore gene conversion events may lead to biased conclusions.

Variations in the length of the middle A-rich region and oligo-dA-rich tails of Alu elements are not uncommon (Economou et al. 1990; Arcot et al. 1995b; Jurka and Pethiyagoda 1995). Microsatellite repeats have been found to be associated with the 3' oligo (dA) tails and the middle A-rich region of Alu elements. In the case of Friedreich ataxia, the most common mutation is the hyperexpansion of a GAA trinucleotide repeat within the middle A-rich region of an Sx Alu (Montermini et al. 1997). However, microsatellites in the middle of Alu elements are not as common because of the much shorter initial length of the middle A-rich region. Arcot et al. (1995b) reported previously that only about one-fourth of the Alu elements containing (AC)_n repeats had them as a part of their middle A-rich region. The one specific example they studied in detail had an evolutionary expansion of the A-rich region (orangutan and gibbon) before the genesis of the AC repeat; suggesting the requirement for an initial expansion. Interestingly, our large-scale analysis of the middle A-rich regions of Ya5 elements demonstrates a trend toward expansion of the A region, providing additional support for this region of the Alu elements to act as a potential nucleus for the genesis of simple sequence repeats.

From our subset of 269 AluYa5 elements, we were able to identify a new Alu subfamily termed Ya5a2. The estimated average age of 0.62 million years (0.28–1.08 million years with 95% confidence) makes Ya5a2 the youngest subfamily of Alu repeats identified in the human genome to date. It is as abundant as the Ya8 subfamily (Roy et al. 1999) and its higher level of insertion polymorphism suggests a higher level of current retroposition. The Ya5a2 subfamily may have originated from a Ya5 Alu element that inserted in a genomic region that favored transcription and corresponding retroposition activity of the element, thereby generating a source gene. The subsequent accumulation of the two specific mutations facilitated the differentiation of the copies made by the Ya5a2 source gene from the larger background of several hundred genomic Ya5 Alu family members. As new Alu elements integrate into the genome in favorable genomic locations, they can occasionally remain retropositionally competent and generate copies of themselves. However, the frequency

of fortuitous insertions of new Alu elements into favorable genomic locations for subsequent mobilization is still a rare event because the continuity of the hierarchical subfamily sequence structure of the Alu elements is largely conserved throughout primate evolution.

Alu elements that are polymorphic for insertion presence/absence have been proven previously to be useful for the study of human population genetics and forensics (Batzer et al. 1991; Jorde et al. 2000; Perna et al. 1992; Batzer et al. 1994; Tishkoff et al. 1996; Stoneking et al. 1997). The identification of a very young Alu subfamily with a high proportion of polymorphic members provides a new source of Alu insertion polymorphisms for the study of human population genetics. However, it is important to note that the Ya5a2 subfamily is extremely small (~35 copies total in a background of >1,000,000) comparable with Ya8, so that an exhaustive analysis of a single human genome would only generate ~20 polymorphic Ya5a2 elements.

Because our analysis of Alu elements related to the Apert's insertion only included ~40% of the human genome (both finished and draft sequence included), there are possibly one or two other perfect complements in the human genome that have not yet been sequenced and may be the actual source gene for these elements. The transcriptional potential of this element would be consistent with its role as the potential source Alu gene. This confirms the existence of minor active source genes that differ from the source gene that generated almost all of the Alu elements present in the human genome today. In addition, the de novo Apert's Alu element was also transcriptionally active. There are two possible explanations for this result. First, the transcriptional capacity of the elements was evaluated by transient transfections in tissue culture. This system does not reflect the influence of chromatin structure and methylation patterns (position effects) on the transcription and presumably retroposition potential of the two Alu repeats. Alternatively, the de novo Apert's Alu element may have inserted in a region of the *FGFR2* gene that fortuitously enhances its own transcription capability. Although further studies will be required to make more definitive statements in this regard, the transcriptional capability of Ya5NBC237 is consistent with one of the many requirements a source gene possesses, making it a plausible candidate source gene for the de novo Apert's insertion.

In summary, the computational analyses of a subset of recently integrated Alu elements demonstrate that Alu sequence evolution is affected by a number of dynamic events. New retroposition-competent Alu source genes, gene conversion, and genetic instability each play an important role in Alu sequence evolution and proliferation within the human genome.

METHODS

Computational Analyses

Screening of the GenBank nr, the htgs, and the gss databases were performed by use of the Advanced Basic Local Alignment Search Tool 2.0 (BLAST) (Altschul et al. 1990) available from the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). For the Ya5 subfamily analysis, the database was searched for matches to the 281 bases of the Ya5 consensus sequence with the following advanced options: -e 1.0 e-120, -b 1000, and -v 1000. A region composed of 500 bases of flanking DNA sequence directly adjacent to the sequences identified from the databases that matched the initial GenBank BLAST query were subjected to annotation by use of either RepeatMasker2 from the University of Washington Genome Center server (<http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>) or Censor from the Genetic Information Research Institute (http://www.girinst.org/Censor_Server/Data_Entry_Forms.html) (Jurka et al. 1996). These programs annotate the repeat sequence content of DNA sequences from humans and rodents. The sequences were then subjected to more detailed analysis by use of MegAlign (DNAStar version 3.1.7 for Windows 3.2). The following parameters were used to select the Ya5 elements to be analyzed: (1) Ya5 had to have all five diagnostic nucleotides (except for the first position, as it is a highly mutable CpG). (2) No truncated Alu elements were included in the analysis. (3) No Alu elements identified as a result of directed cloning strategies designed to identify Alu repeats were included (only those randomly found within larger data sequence). (4) Duplicate Alu elements were eliminated on the basis of flanking sequences. The consensus sequences of the Yb8 and Ya8 subfamilies were used for parallel searches of the three GenBank databases mentioned above. A complete list of the Alu elements identified from the GenBank search is available from M.A.B. or P.L.D. and at <http://www.genome.org/cgi/doi/10.1101/gr152300>.

To search for putative source genes of the Alu elements that have been associated previously with different diseases, the three GenBank databases were searched by use of the sequence of each individual repeat to identify exact complements (Deininger and Batzer 1999).

DNA Samples

Human DNA samples from the European, African-American, Egyptian, and Greenland native population groups were isolated from peripheral blood lymphocytes (Ausubel et al. 1996) that were available from previous studies (Roy et al. 1999).

Oligonucleotide Primer Design and PCR Amplification

A region composed of ~500 bases of flanking unique DNA sequences adjacent to each Alu repeat were used to design primers for 14 Ya5a2 Alu elements (13 exact matches to consensus, Table 2). PCR primers were designed with the Primer3 software (Whitehead Institute for Biomedical Research) (<http://www.genome.wi.mit.edu/cgi-bin/primer/primer3-www.cgi>). The resultant PCR primers were screened against the GenBank nr database for the presence of repetitive elements by use of the BLAST program, and primers that resided within known repetitive elements were discarded and new primers were designed. PCR amplification was carried out in 25- μ L reactions with 50–100 ng of target DNA, 40 pM of each oligonucleotide primer, 200 μ M dNTPs in 50 mM KCl, 1.5

mM MgCl₂, 10 mM Tris-HCl (pH 8.4), and Taq DNA polymerase (1.25 units) as recommended by the supplier (Life Technologies). Each sample was subjected to the following amplification cycle: an initial denaturation of 2:30 min at 94°C, 1 min of denaturation at 94°C, 1 min at the annealing temperature, 1 min of extension at 72°C, repeated for 32 cycles, followed by a final extension at 72°C for 10 min. Twenty microliters of each sample was fractionated on a 2% agarose gel with 0.25 μ g/ml ethidium bromide. PCR products were directly visualized by UV fluorescence. The human genomic diversity associated with each element was determined by the amplification of 20 individuals from each of 4 populations (African American, Greenland native, European, and Egyptian; 160 total chromosomes). The chromosomal location for elements identified from randomly sequenced large-insert clones was determined by PCR analysis of National Institute of General Medical Sciences (NIGMS) human/rodent somatic cell hybrid mapping panels 1 and 2 (Coriell Institute for Medical Research, Camden, NJ).

Construction of Plasmids

The following constructs were made: p⁴¹⁶Ya5NBC237 (416 bp upstream genomic – Alu – 223 bases downstream); p²⁹⁰Ya5Ap (290 bp upstream genomic – Alu – 293 bases); and p^{NP}Ya5NBC237 (no upstream vector flank-Alu – 223 bases). Unless otherwise noted, PCR was performed in 20- μ L reactions by use of an MJ Research PTC 200 thermal cycler with the following conditions: 1X Promega buffer, 1.5 mM MgCl₂, 200 μ M dNTPs, 0.25 μ M primers, 1.5 units of Taq polymerase (Promega) at 94°C for 2 min; 94°C for 20 sec, 55°C (annealing temperature) for 20 sec, 72°C for 1 min, for 30 cycles; 72°C for 3 min. To PCR amplify and clone the 864-bp fragment containing the de novo Alu Ya5 from Apert syndrome patient 1 (accession no. AF097344), the following primers were used: forward, 5'-GGTGTGGCAAAGTGGAGGATGTGTAC-3' and reverse, 5'-TTATTCAAGGATAAAAGGGCATTTC-3' with an annealing temperature of 50°C; and for the 920-bp fragment containing AluYa5NBC237 (accession no. AL031274) the primers used were: forward, 5'-TTATTCCATTG GTCCTTCCACCAG-3' and reverse, 5'-CAGGCAGGGAGG TACTTGCTCTTG-3' with an annealing temperature of 55°C.

For the p^{NP}Ya5NBC237, PCR amplification from the clone was done with the same reverse primer and the FAlu5 primer 5'-GGCCGGCGCGGTGGCTCA-3'.

The final PCR product of the complete construct was cloned into pGEMTeasy Vector System I (Promega). Constructs were subjected to DNA sequence analysis to verify their sequence context. Purified plasmids from the constructs were prepared by alkaline lysis of bacterial cells followed by banding in a CsCl gradient twice. DNA concentrations were determined spectrophotometrically by use of A_{260} and verified by visual examination of ethidium bromide-stained agarose gels.

Alu Transcription in Cell Lines and RNA Analysis

Transient transfections were carried out in the rodent cell line C6 glioma (ATCC CCL107). Monolayers were grown to 50–70% confluence and transfected with 3 μ g of the construct-containing plasmid and 1 μ g of control plasmid (p^{7S1}BC1) by use of LipofectAmine Plus (GIBCO Life Sciences) following the manufacturer's recommended protocol. Total RNA was isolated 16–20 h post-transfection.

RNA was extracted from cell lines utilizing the Trizol Reagent (Life Technologies, Inc.) according to the manufactur-

er's protocol. Equal amounts of RNA were fractionated on a 2% agarose-formaldehyde gel and then transferred to a nylon membrane, Hybond-N (Amersham). Northern blots were hybridized utilizing the following end-labeled oligonucleotide probes: unique-1 5'-TGTGTGTGCCAGTTACCTG-3' (complementary to the 3' end of the control plasmid) and AluYAS-1 5'-ACCGTTAGCCGGGAATGGTC-3' (complementary to Ya5 Alu RNA, but not to 7SL) in 5× SSC, 5× Denhardt's, 1% SDS, and 100 µg/mL herring sperm DNA. Oligonucleotides were end labeled by incorporating [γ -³²P]ATP (Amersham) with T4 polynucleotide kinase (New England BioLabs), and subsequently separated from free label by filtration through a Sephadex G-50 column. Blots were washed three times at 45°C with a low stringency buffer (2× SSC and 1% SDS) and subjected to autoradiography or quantified with a FujiFilm FLA-2000 fluorescent image analyzer (Fuji Photo Film Co. LTD). Statistical analysis was performed with the Jandel SigmaStat Statistical Software Version 2, (Jandel Corporation).

ACKNOWLEDGMENTS

A.M.R. was supported by a Brown Foundation fellowship from the Tulane Cancer Center. This research was supported by National Institutes of Health RO1 GM45668 to P.L.D., Department of the Army DAMD17-98-1-8119 to P.L.D. and M.A.B., and award no. 1999-IJ-CX-K009 from the Office of Justice Programs, National Institute of Justice, Department of Justice to M.A.B. Opinions expressed in this document are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.

Arcot, S.S., Shaikh, T.H., Kim, J., Bennett, L., Alegria-Hartman, M., Nelson, D.O., Deininger, P.L., and Batzer, M.A. 1995a. Sequence diversity and chromosomal distribution of "young" Alu repeats. *Gene* **163**: 273-278.

Arcot, S.S., Wang, Z., Weber, J., Deininger, P.L., and Batzer, M.A. 1995b. Alu repeats: A source for the genesis of primate microsatellites. *Genomics* **29**: 136-144.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. 1996. *Current protocols in molecular biology*, John Wiley & Sons, Inc. Canada.

Batzer, M.A., Kilroy, G., Richard, P.E., Shaikh, T.H., Dessel, T., Hoppens, C., and Deininger, P.L. 1990. Structure and variability of recently inserted Alu family members. *Nucleic Acids Res.* **18**: 6793-6798.

Batzer, M.A., Gudi, V., Mena, J.C., Foltz, D.W., Herrera, R.J., and Deininger, P.L. 1991. Amplification dynamics of human-specific (HS) Alu family members. *Nucleic Acids Res.* **19**: 3619-3623.

Batzer, M.A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D.H., Shaikh, T.H., Novick, G.E., Ioannou, P.A., Scheer, W.D., Herrera, R.J., and Deininger, P.L. 1994. African origin of human-specific polymorphic Alu insertions. *Proc. Natl. Acad. Sci.* **91**: 12288-12292.

Batzer, M.A., Rubin, C.M., Hellmann-Blumberg, U., Alegria-Hartman, M., Leeflang, E.P., Stern, J.D., Bazan, H.A., Shaikh, T.H., Deininger, P.L., and Schmid, C.W. 1995. Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *J. Mol. Biol.* **247**: 418-427.

Batzer, M.A., Deininger, P.L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C.M., Schmid, C.W., Zietkiewicz, E., and Zuckerkandl, E. 1996a. Standardized nomenclature for Alu repeats. *J. Mol. Evol.* **42**: 3-6.

Batzer, M.A., Arcot, S.S., Phinney, J.W., Alegria-Hartman, M., Kass, D.H., Milligan, S.M., Kimpton, C., Gill, P., Hochmeister, M., Ioannou, P.A., et al. 1996b. Genetic variation of recent Alu insertion in human populations. *J. Mol. Evol.* **42**: 22-29.

Belmaaza, A., Wallenburg, J.C., Brouillet, S., Gusew, N., and Chartrand, P. 1990. Genetic exchange between endogenous and exogenous LINE-1 repetitive elements in mouse cells. *Nucleic Acids Res.* **18**: 6385-6391.

Bird, A.P. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**: 1499-1504.

Deininger, P.L. and Daniels, G. 1986. The recent evolution of mammalian repetitive DNA elements. *Trends Genet.* **2**: 76-80.

Deininger, P.L. and Batzer, M.A. 1993. Evolution of retroposons. In *Evolutionary biology* (ed. M.K. Heckht), pp. 157-196. Plenum Publishing, New York, NY.

—. 1999. Alu repeats and human disease. *Mol. Genet. Metab.* **67**: 183-193.

Deininger, P.L., Batzer, M.A., Hutchison, I.C., and Edgell, M. 1992. Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**: 307-312.

Dombroski, B.A., Scott, A.F., and Kazazian, Jr., H.H. 1993. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc. Natl. Acad. Sci.* **90**: 6513-6517.

Economou, E.P., Bergen, A.W., Warren, A.C., and Antonarakis, S.E. 1990. The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. *Proc. Natl. Acad. Sci.* **87**: 2951-2954.

Hoff, E.F., Levin, H.L., and Boeke, J.D. 1998. *Schizosaccharomyces pombe* retrotransposon Tf2 mobilizes primarily through homologous cDNA recombination. *Mol. Cell. Biol.* **18**: 6839-6852.

Jelinek, W.R. and Schmid, C.W. 1982. Repetitive sequences in eukaryotic DNA and their expression. *Annu. Rev. Biochem.* **51**: 813-844.

Jurka, J. and Smith, T. 1988. A fundamental division in the Alu family of repeated sequences. *Proc. Natl. Acad. Sci.* **85**: 4775-4778.

Jurka, J. and Pethiyagoda, C. 1995. Simple repetitive DNA sequences from primates: Compilation and analysis. *J. Mol. Evol.* **40**: 120-126.

Jurka, J., Kaplan, D.J., Duncan, C.H., Walichiewicz, J., Milosavljevic, A., Murali, G., and Solus, J.F. 1993. Identification and characterization of new human medium reiteration frequency repeats. *Nucleic Acids Res.* **21**: 1273-1279.

Jurka, J., Klonowski, P., Dagman, V., and Pelton, P. 1996. CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20**: 119-121.

Jorde, L.B., Watkins, W.S., Bamshad, M.J., Dixon, M.E., Ricker, C.E., Seielstad, M.T., and Batzer, M.A. 2000. The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data. *Am. J. Hum. Genet.* **66**: 979-988.

Kapitonov, V. and Jurka, J. 1996. The age of Alu subfamilies. *J. Mol. Evol.* **42**: 59-65.

Kass, D.H., Batzer, M.A., and Deininger, P.L. 1995. Gene conversion as a secondary mechanism in SINE evolution. *Mol. Cell. Biol.* **15**: 19-25.

Labuda, D. and Striker, G. 1989. Sequence conservation in Alu evolution. *Nucleic Acids Res.* **17**: 2477-2491.

Lester, T., McMahon, C., VanRegemorter, N., Jones, A., and Genet, S. 1997. X-linked immunodeficiency caused by insertion of Alu repeat sequences. *J. Med. Gen. Suppl.* **34**: S81.

Miki, Y., Katagiri, T., Kasumi, F., Yoshimoto, T., and Nakamura, Y. 1996. Mutation analysis in the BRCA2 gene in primary breast cancers. *Nat. Genet.* **13**: 245-247.

Miyamoto, M.M., Slightom, J.L., and Goodman, M. 1987. Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. *Science* **238**: 369–373.

Montermini, L., Andermann, E., Labuda, M., Richter, A., Pandolfo, M., Cavalcanti, F., Pianese, L., Iodice, L., Farina, G., Monticelli, A., et al. 1997. The Friedreich ataxia GAA triplet repeat: Premutation and normal alleles. *Hum. Mol. Genet.* **6**: 1261–1266.

Oldridge, M., Zackai, E.H., McDonald-McGinn, D.M., Iseki, S., Morris-Kay, G.M., Twigg, S.R., Johnson, D., Wall, S.A., Jiang, W., et al. 1999. De novo Alu-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome. *Am. J. Hum. Genet.* **64**: 446–461.

Perna, N.T., Batzer, M.A., Deininger, P.L., and Stoneking, M. 1992. Alu insertion polymorphism: A new type of marker for human population studies. *Hum. Biol.* **64**: 641–648.

Rogers, J.R. and Willison, K.R. 1983. A major rearrangement in the H-2 complex of mouse t haplotypes. *Nature* **304**: 549–552.

Roy, A.M., Carroll, M.L., Kass, D.H., Nguyen, S.V., Salem, A-H., Batzer, M.A., and Deininger, P.L. 1999. Recently integrated human Alu repeats: Finding needles in the haystack. *Genetica* **107**: 149–161.

Schmid, C.W. and Maraia, R. 1992. Transcriptional regulation and transpositional selection of active SINE sequences. *Curr. Opin. Genet. Dev.* **2**: 874–882.

Shaikh, T.H. and Deininger, P.L. 1996. The role and amplification of the HS Alu subfamily founder gene. *J. Mol. Evol.* **42**: 15–21.

Shaikh, T.H., Roy, A.M., Kim, J., Batzer, M.A., and Deininger, P.L. 1997. cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts. *J. Mol. Biol.* **271**: 222–234.

Shen, M., Batzer, M.A., and Deininger, P.L. 1991. Evolution of the master Alu gene(s). *J. Mol. Evol.* **33**: 311–320.

Sinnett, D., Richer, C., Deragon, J.M., and Labuda, D. 1992. Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences. *J. Mol. Biol.* **226**: 689–706.

Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H., and Deininger, P.L. 1987. Clustering and sub-family relationships of the Alu family in the human genome. *Mol. Biol. Evol.* **4**: 19–29.

Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**: 657–663.

Stoneking, M., Fontius, J.J., Clifford, S.L., Soodyall, H., Arcot, S.S., Saha, N., Jenkins, T., Tahir, M.A., Deininger, P.L., and Batzer, M.A. 1997. Alu insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. *Genome Res.* **7**: 1061–1071.

Tishkoff, S.A., Ruano, G., Kidd, J.R., and Kidd, K.K. 1996. Distribution and frequency of a polymorphic Alu insertion at the plasminogen activator locus in humans. *Hum. Genet.* **97**: 759–764.

Wallace, M.R., Andersen, L.B., Saulino, A.M., Gregory, P.E., Glover, T.W., and Collins, F.S. 1991. A de novo Alu insertion results in neurofibromatosis type 1. *Nature* **353**: 864–866.

Weiner, A., Deininger, P.L., and Efstradiatis, A. 1986. The Reverse flow of genetic information: Pseudogenes and transposable elements derived from nonviral cellular RNA. *Annu. Rev. Biochem.* **55**: 631–661.

Willard, C., Nguyen, H.T., and Schmid, C.W. 1987. Existence of at least three distinct Alu subfamilies. *J. Mol. Evol.* **26**: 180–186.

Received June 14, 2000; accepted in revised form August 9, 2000.

MINIREVIEW

Alu Repeats and Human Disease

Prescott L. Deininger*†¹ and Mark A. Batzer‡

*Tulane Cancer Center, SL-66, and Department of Environmental Health Sciences, Tulane University Medical Center, 1430 Tulane Avenue, New Orleans, Louisiana 70112; †Laboratory of Molecular Genetics, Ochsner Medical Foundation, 1516 Jefferson Highway, New Orleans, Louisiana 70121; and ‡Departments of Pathology, Biochemistry and Molecular Biology, Biometry and Genetics, Stanley S. Scott Cancer Center, Neuroscience Center of Excellence, Louisiana State University Medical Center, 1901 Perdido Street, New Orleans, Louisiana 70112

Received March 25, 1999, and in revised form March 31, 1999

Alu elements have amplified in primate genomes through a RNA-dependent mechanism, termed retroposition, and have reached a copy number in excess of 500,000 copies per human genome. These elements have been proposed to have a number of functions in the human genome, and have certainly had a major impact on genomic architecture. Alu elements continue to amplify at a rate of about one insertion every 200 new births. We have found 16 examples of diseases caused by the insertion of Alu elements, suggesting that they may contribute to about 0.1% of human genetic disorders by this mechanism. The large number of Alu elements within primate genomes also provides abundant opportunities for unequal homologous recombination events. These events often occur intrachromosomally, resulting in deletion or duplication of exons in a gene, but they also can occur interchromosomally, causing more complex chromosomal abnormalities. We have found 33 cases of germline genetic diseases and 16 cases of cancer caused by unequal homologous recombination between Alu repeats. We estimate that this mode of mutagenesis accounts for another 0.3% of human genetic diseases. Between these different mechanisms, Alu elements have not only contributed a great deal to the evolution of the genome but also continue to contribute to a significant portion of human genetic diseases. © 1999 Academic Press

¹ To whom correspondence should be addressed at Tulane Cancer Center, SL-66, Tulane University Medical Center, 1430 Tulane Ave., New Orleans, LA 70112. Fax: (504) 588-5516. E-mail: pdeinin@tcs.tulane.edu.

Key Words: Alu repeats; recombination; insertion mutation; human disease; genetic diversity.

THE SPREAD OF Alu ELEMENTS IN THE HUMAN GENOME

Alu elements represent a sequence of approximately 300 nucleotides (nt) in length that are transcribed by RNA polymerase III. The RNA transcript is then reverse-transcribed and inserted into a new location in the genome. This RNA-mediated process for making new copies of the element is termed retroposition (1). Different Alu elements in the genome are not identical to one another. It appears that Alu elements that have integrated recently within the genome are quite homogeneous, and almost exact copies of one another (2). However, the older copies have accumulated random mutations, making them typically divergent by 20% or more from one another at the sequence level (3).

Alu elements began inserting early in primate evolution, approximately 65 mya (3). Although there are some related elements in mammals outside of the primate order, they do not have the specific structure of Alu elements. The rate of Alu amplification appears to have reached a maximum between 35 and 60 mya, and is currently amplifying at only 1% of the maximum rate. There are probably only about 2000 Alus specific to the human genome, and not found in chimpanzee and gorilla. Thus, about 99.8% of the 500,000 Alus in the human genome can



TABLE 1
Alu Insertions and Disease

Locus	Distribution	Subfamily	Disease	Reference
CaR	Familial	Ya4	Hypocalciuric hypercalcemia and neonatal severe hyperparathyroidism	(51)
Mlvi-2	<i>De novo</i> (somatic?)	Ya5	Associated with leukemia	(52)
NF1	<i>De novo</i>	Ya5	Neurofibromatosis	(53)
PROGINS	About 50%	Ya5	Linked with ovarian carcinoma	(54)
IL2RG	Familial	Ya5	XSCID	(55)
ACE	About 50%	Ya5	Linked with protection from heart disease	(35)
Factor IX	A grandparent	Ya5	Hemophilia	(56)
EYA1	<i>De novo</i>	Ya5	Branchio-oto-renal syndrome	(57)
2 × FGFR2	<i>De novo</i>	Ya5 & Yb8	Apert's syndrome	(41)
Cholinesterase	One Japanese family	Yb8	Cholinesterase deficiency	(58)
APC	Familial	Yb8	Hereditary desmoid disease	(59)
Btk	Familial	Y	X-linked agammaglobulinaemia	(55)
C1 inhibitor	<i>De novo</i>	Y	Complement deficiency	(60)
BRCA2	<i>De novo</i>	Y	Breast cancer	(61)
GK	?	Y	Glycerol kinase deficiency	(62)

be found at the same locus in all of the great apes, and 85% of the elements at specific loci can be found in all monkeys. Our best estimates of Alu amplification in the human genome are that there is one new insert in about every 200 new births (4). Although this is well below the peak rate, it is still high enough to represent a significant factor in human mutagenesis.

In addition to random mutations, which occur to Alu elements after their insertion in the genome, there are specific base changes that allow separation of Alu elements into different subfamilies (5–10). The different subfamilies were all inserted at different stages of primate evolution. Almost all of the insertions that have occurred specifically in the human genome come from four closely related subfamilies, Alu Y, Ya5, Ya8, and Yb8. Ya5 and Yb8 inserts represent the majority of the inserts and Alu Y inserts are relatively rare. All of the new inserts belong to a small group of the most recently created subfamilies (see Table 1). This demonstrates that only a small subset of Alus is capable of amplification (11).

Several explanations for the selective amplification of specific subfamilies have been proposed. One likely explanation is that a few specific loci are capable of active amplification, while almost all other loci are not, and that there are almost no such loci in the older subfamilies (11). Alternatively, one has to propose that loci from all subfamilies express, but that the RNAs expressed from the newer subfami-

lies interact with the retroposition apparatus much better than the older subfamily RNAs (12,13).

Alus AND L1 ELEMENTS

The other major mobile element in the human genome is the L1 element. Alu elements are RNA polymerase III-derived transcripts that have no coding capacity. Thus, they do not code for any proteins that might be involved in the retroposition process. L1 repeats, on the other hand, are much longer and have two open-reading frames (reviewed in (14)). One open-reading frame apparently codes for an RNA-binding protein whose exact function is unknown. The other open-reading frame codes for a protein that includes domains for reverse transcriptase, as well as for an endonuclease that apparently nicks the genome at the site of insertion (15–17). An assay that allows rapid L1 retroposition in cultured cells has been devised recently (18). This assay facilitates the dissection of the details of the L1 retroposition mechanism.

Alu elements must obtain the enzymes for their retroposition from somewhere. In addition, there are striking similarities between the mechanisms of Alu and L1 retroposition that make it very attractive to think that L1 elements may supply the necessary components for Alu retroposition (15,16,19,20). This idea is certainly very attractive, and thus the rate of Alu retroposition may be very dependent on the rate and evolution of L1 elements.

Alu ELEMENTS: FUNCTIONAL ROLE OR A PARASITE'S PARASITE

Alu repeats represent over 5% of the mass of the human genome. They are also spread throughout the entire genome, at varying densities. These observations, along with other specific properties of the Alu elements have led to a number of hypothetical functions for the Alu elements that might explain their ubiquitous presence in primate genomes. Some of the proposed roles involve an everyday function for the cell, while others are of a more sporadic nature.

The first role ever proposed for Alu elements was that they might be origins of DNA replication (21). This role is consistent with their high copy number and dispersed nature, but has not been substantiated by direct experimentation and seems like too important a function to be served by an element that is not found outside of primates.

More recently, evidence has been presented that Alu RNAs may stimulate protein translation by inhibiting a RNA-dependent protein kinase, PKR (22-24). Because Alu RNAs from many loci are stimulated by a number of cellular stresses, such as viral infection and heat shock, this would provide a mechanism by which dispersed sequences may contribute to a cellular process as a group. If this is a function of Alu elements, then it is likely to represent only a slightly modified regulation seen in nonprimate species that is filled by other RNAs or molecules in those species.

Evidence has been presented in yeast that retrotransposable elements may aid in healing chromosomal breaks (25,26). This suggests the possibility that Alu and L1 elements may provide the same role in the human genome.

There are several thoughts concerning the possible roles of Alu elements in the evolution of the human genome. As discussed below, Alu elements can lead to unequal recombination that results in deletion or duplication of sequences. These events could allow duplication of exons and therefore formation of new protein variants. They can also contribute to interchromosomal recombination that may lead to cytogenetic alterations that are involved in human speciation.

There are also several ways in which Alu repeats have been proposed to influence the evolution of gene expression. Because Alu elements are rich in CpG dinucleotides that represent the substrate for genomic methylation, Alu elements rep-

resent CpG-rich islands that make up about 30% of the methylation sites in the human genome (24). When an Alu element inserts in a new location in the genome, it introduces a CpG island at that new location. CpG islands have been associated with gene regulation, as well as imprinting of genes, and therefore Alu elements may contribute to the evolution of gene expression and imprinting in the human genome. In addition, Alu elements have been found to carry functional promoter elements for several of the steroid hormone receptors (27,28). Thus, insertion of a new Alu element in the vicinity of a gene may introduce new transcription factor-binding sites that could alter the regulation of gene expression. There are a number of cases where elements that influence gene expression have been mapped to within an Alu repeat (29), demonstrating that the introduction of these sequences can at least occasionally contribute to gene expression and regulation.

Although, there are numerous cases where individual Alu elements have had a positive impact on the human genome, it might be argued that none of them has been confirmed as a function. In this sense we would not define something that happens in a positive sense every few thousand years as being a function, because it would be occurring too sporadically to apply a positive selection for the presence of Alu elements. In addition, studies of individual Alu elements demonstrate that there is essentially no selective pressure on any given Alu repeat, although it is possible that selection does exist for a handful of master elements. Thus, it has been argued that Alu and L1 elements may both represent "selfish" DNA, or DNA that is only working to replicate itself. Selfish DNA may often have negative impacts on the host, but can be tolerated if it does not have too strong an adverse affect. Selfish DNA may also occasionally have positive benefits, but only by chance, and not by functional design. If L1 elements are essentially a parasite within the human genome, and if Alu relies on L1 elements for their amplification process, then one might describe Alu as a "parasite's parasite."

Alus AS MARKERS FOR HUMAN DIVERSITY

Although there is still a question as to whether there is a true functional role for Alu elements in the human genome, Alu elements have proved to be

useful in studies of human DNA. The presence of Alu repeats located ubiquitously throughout the human genome, but not in nonprimate species, has allowed detection of human DNA sequences that have been transfected into the cells of other organisms, such as mice. This has been useful in marker-rescue experiments in isolating a number of genes, including the first examples of oncogenes isolated by transforming rodent cell lines with human tumor DNAs (30). More recently, inter-Alu PCR (31,32) has found a broad range of uses in isolating specific human DNA regions from mouse/human hybrid cell lines and other complex sources containing large segments of human DNA.

Recent Alu insertions have also proven useful in a number of human population studies. In particular, there are over 1000 Alu insertions that occurred recently enough to be present only in a subset of human chromosomes. Because there does not seem to be any specific mechanism for removing Alu elements from the genome, once inserted they make a very stable genetic marker (33,34). This observation, along with the extremely low probability that any two recently integrated elements have inserted independently in the same chromosomal location, makes Alu insertions one of the best identical-by-descent (IBD) markers for human evolution studies. Any two individuals sharing an Alu insert almost certainly do so because they share a common ancestor in which the insertion occurred. Table 1 includes an example of an Alu insertion in the angiotensin-converting enzyme (ACE) locus that shows a useful association with protective advantages from heart disease (35). Many other Alu insertion polymorphisms have been identified either in random genomic loci or in specific genes, but without any known disease association. These Alu insertions are easy to assay for their presence or absence in a chromosomal location and have been found to be very powerful markers for human forensic and molecular anthropology studies (36,37).

RETROPOSITION OF Alu ELEMENTS AND DISEASE

Alu elements are located throughout the genome and in almost any location within a gene except those in which they would totally disrupt the function of that gene. Figure 1 illustrates some of the positions relative to a typical gene structure in which Alu may land. Alus landing far enough upstream of a gene may have no influence on that

gene's expression. However, Alus landing in or near the promoter/enhancer regions of a gene have been found to influence the expression of specific genes (reviewed in (29)), as well as to have the general potential to add transcription elements, like steroid hormone receptor elements (27,28), to the upstream gene region.

Very few Alu elements are found within the 5' noncoding or coding regions of exons, presumably because insertions in those locations are too disruptive to gene function. There are a number of instances where Alu elements have been found to be part of the region coding for the carboxy-terminus of a protein product (38,39). Presumably these Alus insert far enough downstream in the coding sequence to result in a new carboxy-terminus that does not disrupt the structure of the protein.

Insertions into the 3' noncoding regions of genes are found commonly and appear to have few negative affects. Similarly Alus are commonly found in introns, demonstrating that Alu insertions in much of the intronic region do not alter gene function significantly.

The vast majority of Alu insertions that have led to human disease insert into coding exons, or into introns relatively near an exon and presumably alter splicing. Table 1 is a list of the genetic defects that are thought to be caused by Alu insertion events. Not all of these cases have been demonstrated to be directly causative for the disease, but the rarity of Alu insertion events, coupled with the lack of other detectable mutations in these cases, strongly indicates that these are the causative events. The ACE insertion (35,40) is likely to be one example, however, that shows association with disease, but is highly unlikely to be the causative event.

The above examples demonstrate that Alu insertions are capable of causing genetic defects which lead to human disease. Examples of this type are being found at an increasing frequency as the tools for genetic analysis allow more mutations to be detected. Finding 16 Alu-based insertion mutations in the Human Genetic Mutation Database that contains 14374 characterized human mutations suggests that Alu elements contribute to approximately 0.1% of human genetic diseases. This number agrees well with a previous calculation based on a similar dataset of mutations where Alu and L1 insertions were estimated to each contribute approximately 0.075% of human mutations (16). In some cases, the insertional mutagenesis may make detection of mutations easier, biasing the results in favor of the

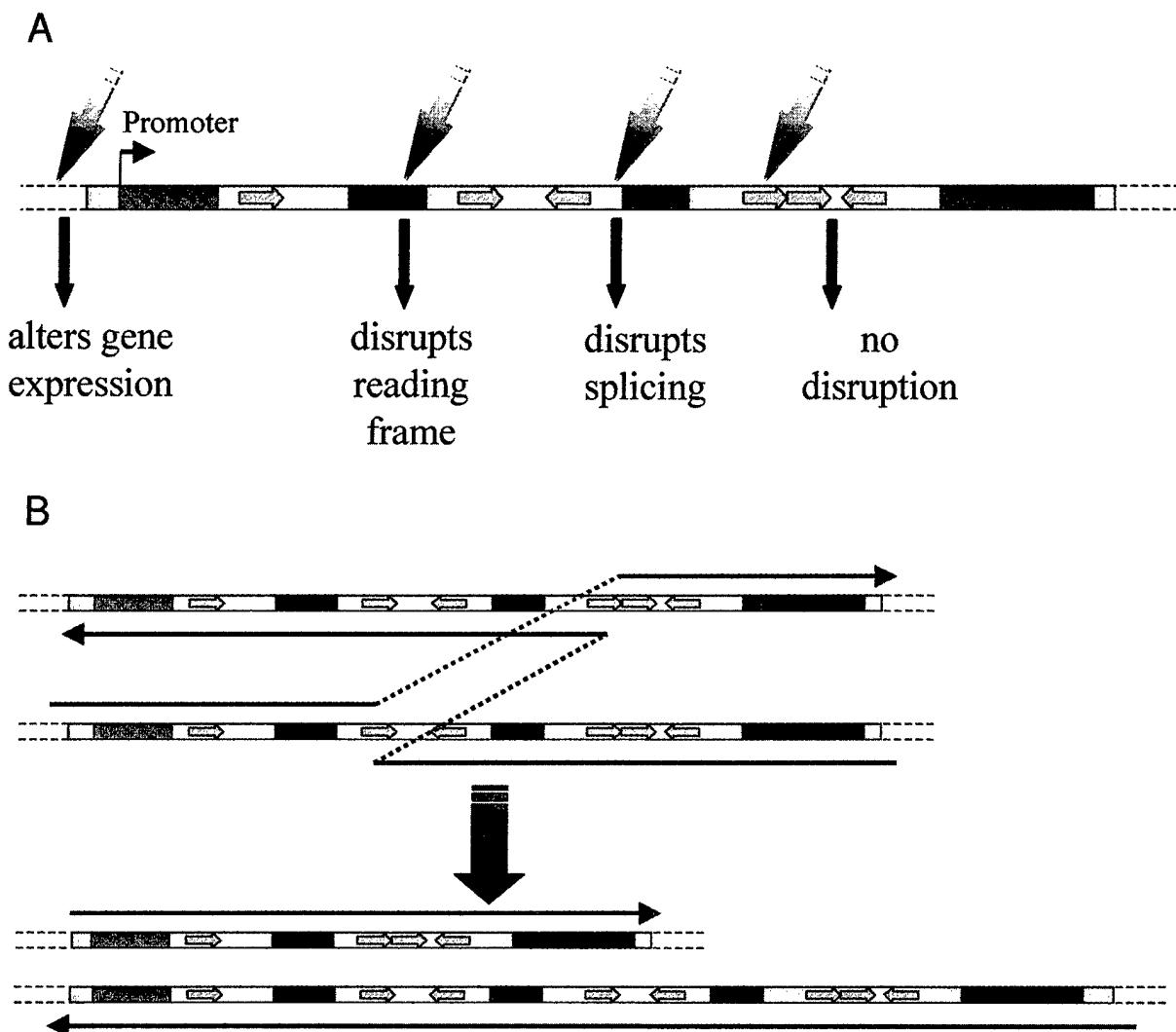


FIG. 1. Schematic of Alu-induced damage to the human genome. Panel A illustrates some of the potential consequences of insertion of a new element in the vicinity of a gene. The colored boxes represent various exons of the gene. The red arrows show existing Alu elements oriented in different directions in the introns of the gene. Depending on the site of insertion, the Alu element has varied probability of impact on the genome as shown. Panel B illustrates an unequal, homologous recombination occurring between two Alu elements in different introns of a gene. The arrows broken by dotted lines show the path of the recombination event. The genes below show that one copy will have a deletion while the other will duplicate gene sequences. Either is likely to be deleterious.

detection of Alu insertions. However, many mutation detection strategies are designed to identify point mutations, particularly in coding regions, and may overlook insertions, particularly if they occur in introns. In addition, many new mobile element insertions may be lethal during embryogenesis. Therefore, it is likely that these estimates of insertion frequencies are underestimates of the true contribution of new Alu insertions to human disease.

We expect that with increasing study of mutations, it will be found that some genetic diseases are

more likely than others to result from retroposon insertion. It has certainly been observed that some genes have a much higher Alu repeat content, making it reasonable that they will have a higher frequency of disabling Alu insertions. It has been observed that 2 out of 258 mutations in the FGFR2 gene were caused by Alu insertions (41). This is the first case of multiple Alu insertion mutations being detected associated with a single disease, suggesting that this genetic locus may be more susceptible to retroposon insertions than other regions of the ge-

TABLE 2
Alu/Alu Recombination and Germ-Line Disease

Locus	Distribution	Disease	Reference
8 × LDLR	Kindreds	Hypercholesterolemia	(63-67)
5 × α -globin	Kindreds	α -thalassaemia	(68-71)
5 × C1 inhibitor	Kindred	Angioneurotic edema	(60,72)
Lys Hydrol.	Kindreds	Ehlers-Danlos syndrome	(73)
DMD	Kindred	Duchenne's muscular dystrophy	(74)
ADA	One patient	ADA deficiency-SCID	(75)
Apo B	One patient	Hypo-betalipoproteinemia	(76)
Ins. Rec. β	One patient	Insulin-independent diabetes	(77)
α -gal A	One patient	Fabry disease	(78)
HPRT	One patient	Lesch-Nyhan syndrome	(79)
Plat. Fibrinogen Receptor	Kindred	Glanzmann thrombasthenia	(80)
Phosphorylase kinase	One patient	Glycogen storage disease	(81)
GALNS	One patient	Mucopolysaccharidosis type IVA	(82)
Antithrombin	One patient	Thrombophilia	(83)
XY	One patient	XX male	(84)
β -HEXA	Classic form of disease	Tay Sachs	(85)
C3	Kindred	C3 deficiency	(86)
HEXB	27% of patients	Sandhoff's disease	(87)

nome. However, the number of insertions found so far is still fairly low making more definitive conclusions difficult.

RECOMBINATION BETWEEN Alu ELEMENTS ASSOCIATED WITH DISEASE

In addition to the potential impact of Alu element insertions in causing human disease, their dispersion throughout the genome provides ample opportunity for unequal homologous recombination which leads to a much higher level of mutations. Figure 1B illustrates how this unequal recombination can cause insertion or deletion mutations. When recombination occurs between Alu elements on the same chromosome, the result is that there is either duplication or deletion of the sequences between the Alus. Recombination may also occur between Alu elements on different chromosomes, resulting in chromosomal translocations or more complex chromosomal rearrangements.

Table 2 presents a compilation of Alu/Alu recombination events that have contributed to germ-line disease with Alu-based recombination events associated with cancer shown in Table 3. There are many more recombination than insertion events contributing to disease and the table of recombination events is not intended to be exhaustive in presenting all of the Alu/Alu recombinations that have contributed to human disease. In addition, there are many

recombination events that occurred between an Alu element and some other non-Alu-related sequence which may have been influenced by the presence of the Alu element (42). Although single Alu elements may contribute specifically to such recombination events, we have made no efforts to collect those data. The mutations resulting from Alu/Alu recombination include 33 mutations that are the result of germ-line recombination and 16 mutations that are the result of somatic events that led to cancer. Based on the calculations in the previous section, the germ-line recombination mutants would represent about 0.3% of mutants characterized. We expect that this number is an underestimate as mutation schemes aimed at detecting point mutants would often be expected to overlook large duplication and deletion events, and we have probably not reported all known Alu/Alu recombinations in the tables.

The data in Tables 2 and 3 show that Alu/Alu recombination events are highly biased towards specific genes. The first to show evidence for this was the LDLR gene, which has at least eight independent cases. It was also reported that these recombination events appeared to take place in a preferred location within the Alu element (42,43). These data suggested that Alu elements may represent hot spots for recombination by a mechanism that was more than simple homologous recombination. Multiple Alu/Alu recombination events have also occurred in the germ line involving two other genes.

TABLE 3
Alu/Alu Recombination and Cancer

Locus	Distribution	Disease	Reference
10 × ALL-1 (MLL)	Somatic	Acute myelogenous leukemia	(88–90)
2 × BRCA1	Somatic and kindreds	Breast cancer	(91,92)
MLH1	Two kindreds	HNPPCC	(93)
TRE	Somatic	Ewing's sarcoma	(94)
RB	Common	Association with glioma	(95)
EWS	Subset of Africans	Protective against Ewing sarcoma?	(96)

Even more striking is the preferential recombination seen in somatic recombination. The All-1 gene which participates in a high proportion of acute leukemias is another hotspot for Alu/Alu recombination. This includes intragenic recombination which is the major cause of acute myelogenous leukemia in individuals without a cytogenetic defect, as well as a possible contribution to recombination between the All-1 gene and other chromosomal loci in causing more complex cytogenetic defects associated with leukemia (44–46).

The genes that show high levels of Alu/Alu recombination tend to have a large number of Alu sequences. Although Alu density may help contribute to this recombination, the correlation does not seem to hold up upon analysis of other Alu-rich genes. Therefore, it seems likely that some other factor contributes to the high recombination rates seen in these genes and that the Alu elements are likely to help in that process rather than to be the primary cause.

It has generally been found that longer stretches of sequence identity allow more efficient homologous recombination and that 300 bp of imperfect sequence identity would represent a relatively inefficient target (47). Therefore, as Alu elements accumulate random mutations after integration in the genome their recombination potential gradually decreases. Thus, early in primate evolution when a high proportion of Alu elements were closer matches to one another, Alu/Alu recombination may have contributed even more to the evolution and reshaping of primate genomes.

Based on the above considerations, one might expect the much longer L1 family of elements to contribute significantly to recombination, as well. Surprisingly, we are familiar with only two L1/L1 recombination events in the human genome (48). Therefore, it would appear that: (1) L1 elements are located in less recombinogenic regions of the human

genome; (2) the approximately 10-fold lower copy number of L1 elements is more than enough to offset their larger size in terms of probabilities of recombination; (3) some basic property of the Alu elements themselves makes them recombinogenic; or (4) the larger average spacing between L1 elements causes the vast majority of L1/L1 recombination events to be lethal. It is possible that all of these factors may contribute to this observed difference. Transient transfection experiments suggest that the third possibility may not be true since Alu sequences did not recombine more frequently than other control sequences (49). However, in their native chromatin environment, or in specific cell types or cell stimuli *in vivo*, Alus may still respond with higher recombination rates. We believe that the fourth possibility may be the dominant factor, however. The vast majority of Alu/Alu recombination events listed in the tables represent recombination between Alu elements within the same gene. This limits the effect of the recombination to a single gene defect. With their lower copy number and tendency to be located between genes rather than in genes, L1/L1 recombination events are likely either to involve only intergenic regions or to involve a much larger region that may cause defects in several genes simultaneously, resulting in loss of viability.

There is growing evidence that repetitive DNAs contribute to disease either through the mutations they cause during the retroposition process that forms them (16,50) or through recombination processes involving unequal cross-overs of repetitive elements. These recombination events may involve repetitive sequences of various repetition frequencies with the likelihood that longer and more perfect repeats that are near one another probably recombine well, while short, mismatched repeats (like Alu) recombine relatively poorly. However, the extremely high copy number of Alu elements makes them a

major factor in the molecular basis of human diseases.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health RO1 GM45668 to P.L.D., Department of the Army DAMD17-98-1-8119 to P.L.D. and M.A.B., and National Science Foundation SBR-9610147 to M.A.B.

REFERENCES

- Rogers J. Retroposons defined. *Nature* **301**:460, 1998.
- Batzer M, Kilroy G, Richard P, Shaikh T, Dessel T, Hoppens C, Deininger P. Structure and variability of recently inserted Alu family members. *Nucleic Acids Res* **18**:6793–6798, 1990.
- Shen M, Batzer M, Deininger P. Evolution of the master Alu gene(s). *J Mol Evol* **33**:311–320, 1991.
- Deininger P, Batzer M. SINE master genes and population biology. In *The Impact of Short, Interspersed Elements (SINEs) on the Host Genome* (Maraia R, Ed.). Georgetown, TX: Landes, pp 43–60, 1995.
- Batzer M, Deininger P. A human-specific subfamily of Alu sequences. *Genomics* **9**:481–487, 1991.
- Jurka J, Smith T. A fundamental division in the Alu family of repeated sequences. *Proc Natl Acad Sci USA* **85**:4775–4779, 1988.
- Matera AG, Hellmann U, Schmid CW. A transpositionally and transcriptionally competent Alu subfamily. *Mol Cell Biol* **10**:5424–5432, 1990.
- Quentin Y. The Alu family developed through successive waves of fixation closely connected with primate lineage history. *J Mol Evol* **27**:194–199, 1988.
- Slagel V, Flemington E, Traina-Dorge V, Bradshaw H, Deininger P. Clustering and sub-family relationships of the Alu family in the human genome. *Mol Biol Evol* **4**:19–29, 1987.
- Willard C, Nguyen HT, Schmid CW. Existence of at least three distinct Alu subfamilies. *J Mol Evol* **26**:180–186, 1987.
- Deininger P, Batzer M, Hutchison C, Edgell M. Master genes in mammalian repetitive DNA amplification. *Trends Genet* **8**:307–312, 1992.
- Sinnett D, Richer C, Deragon JM, Labuda D. Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences. *J Mol Biol* **226**:689–706, 1992.
- Kim J, Kass DH, Deininger PL. Transcription and processing of the rodent ID repeat family in germline and somatic cells. *Nucleic Acids Res* **23**:2245–2251, 1995.
- Boeke JD. LINEs and Alus—The polyA connection. *Nature Genet* **16**:6–7, 1997.
- Feng Q, Moran JV, Kazazian HH Jr, Boeke JD. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**:905–916, 1996.
- Kazazian HH, Jr, Moran JV. The impact of L1 retrotransposons on the human genome. *Nature Genet* **19**:19–24, 1998.
- Mathias SL, Scott AF, Kazazian HH Jr, Boeke JD, Gabriel A. Reverse transcriptase encoded by a human transposable element. *Science* **254**:1808–1810, 1991.
- Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**:917–927, 1996.
- Almenoff JS, Jurka J, Schoolnik GK. Induction of heat-stable enterotoxin receptor activity by a human Alu repeat. *J Biol Chem* **269**:16610–16617, 1994.
- Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci USA* **94**:1872–1877, 1997.
- Jelinek W, Toomey T, Leinwand L, Duncan CH, Biro PA, Choudary PV, Weissman S, Rubin C, Houck C, Deininger PL, Schmid CW. Ubiquitous, interspersed repeated sequences in mammalian genomes. *Proc Natl Acad Sci USA* **77**:1398–1402, 1980.
- Batzer MA, Deininger PL, Hellmann-Blumberg U, Jurka J, Labuda D, Rubin CM, Schmid CW, Zietkiewicz E, Zuckerkandl E. Standardized nomenclature for Alu repeats. *J Mol Evol* **42**:3–6, 1996.
- Chu WM, Ballard R, Carpick BW, Williams BR, Schmid CW. Potential Alu function: Regulation of the activity of double-stranded RNA-activated kinase PKR. *Mol Cell Biol* **18**:58–68, 1998.
- Schmid CW. Does SINE evolution preclude alu function? *Nucleic Acids Res* **26**:4541–4550, 1998.
- Garfinkel DJ. Genetic loose change: How retroelements and reverse transcriptase heal broken chromosomes. *Trends Microbiol* **5**:173–175, 1997.
- Moore JK, Haber JE. Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. *Nature* **383**:644–646, 1996.
- Vansant G, Reynolds WF. The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. *Proc Natl Acad Sci USA* **92**:8229–8233, 1995.
- Norris J, Fan D, Aleman C, Marks JR, Futreal PA, Wiseman RW, Iglehart JD, Deininger PL, McDonnell DP. Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *J Biol Chem* **270**:22777–22782, 1995.
- Britten RJ. DNA sequence insertion and evolutionary variation in gene regulation. *Proc Natl Acad Sci USA* **93**:9374–9377, 1996.
- Shih C, Weinberg RA. Isolation of a transforming sequence from a human bladder carcinoma cell line. *Cell* **29**:161–169, 1982.
- Nelson DL, Ballabio A, Victoria MF, Pieretti M, Bies RD, Gibbs RA, Maley JA, Chinault AC, Webster TD, Caskey CT. Alu-primed polymerase chain reaction for regional assignment of 110 yeast artificial chromosome clones from the human X chromosome: Identification of clones associated with a disease locus. *Proc Natl Acad Sci USA* **88**:6157–6161, 1991.
- Ledbetter SA, Nelson DL, Warren ST, Ledbetter DH. Rapid isolation of DNA probes within specific chromosome regions by interspersed repetitive sequence polymerase chain reaction. *Genomics* **6**:475–481, 1990.

33. Batzer M, Arcot S, Phinney J, Alegria-Hartman M, Kass D, Milligan S, Kimpton C, Gill P, Hochmeister M, Ioannou P, Herrera R, Boudreau D, Scheer WD, Keats B, Deininger P, Stoneking M. Genetic variation of recent *Alu* insertions in human populations. *J Mol Evol* **42**:22–29, 1996.

34. Perna N, Batzer M, Deininger P, Stoneking M. *Alu* insertion polymorphism: A new type of marker for human population studies. *Hum Biol* **64**:641–648, 1992.

35. Cambien F, Poirier O, Lecerf L, Evans A, Cambou JP, Arveiler D, Luc G, Bard JM, Bara L, Ricard S. Deletion polymorphism in the gene for angiotensin-converting enzyme is a potent risk factor for myocardial infarction. *Nature* **359**:641–644, 1992.

36. Batzer M, Stoneking M, Alegria-Hartman M, Bazan H, Kass D, Shaikh T, Novick G, Ioannou PA. African origin of human-specific polymorphic *Alu* insertions. *Proc Natl Acad Sci USA* **91**:12288–12292, 1994.

37. Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, Jenkins T, Tahir MA, Deininger PL, Batzer MA. *Alu* insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. *Genome Res* **7**:1061–1071, 1997.

38. Makalowski W, Mitchell GA, Labuda D. *Alu* sequences in the coding regions of mRNA: A source of protein variability. *Trends Genet* **10**:188–193, 1994.

39. Britten RJ. Mobile elements inserted in the distant past have taken on important functions. *Gene* **205**:177–182, 1997.

40. Tiret L, Rigat B, Visvikis S, Breda C, Corvol P, Cambien F, Soubrier F. Evidence, from combined segregation and linkage analysis, that a variant of the angiotensin I-converting enzyme (ACE) gene controls plasma ACE levels. *Am J Hum Genet* **51**:197–205, 1992.

41. Oldridge M, Zackai EH, McDonald-McGinn DM, Iseki S, Morriess-Kay GM, Twigg SR, Johnson D, Wall SA, Jiang W, Theda C, Jabs EW, Wilkie AO. De novo *Alu*-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome. *Am J Hum Genet* **64**:446–461, 1999.

42. Rudiger NS, Gregersen N, Kielland-Brandt MC. One short well conserved region of *Alu*-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. *Nucleic Acids Res* **23**:256–260, 1995.

43. Lehrman MA, Russell DW, Goldstein JL, Brown MS. *Alu*-*Alu* recombination deletes splice acceptor sites and produces secreted low density lipoprotein receptor in a subject with familial hypercholesterolemia. *J Biol Chem* **262**:3354–3361, 1987.

44. Jeffs AR, Benjes SM, Smith TL, Sowerby SJ, Morris CM. The BCR gene recombines preferentially with *Alu* elements in complex BCR-ABL translocations of chronic myeloid leukaemia. *Hum Mol Genet* **7**:767–776, 1998.

45. Chen SJ, Chen Z, Font MP, d'Auriol L, Larsen CJ, Berger R. Structural alterations of the BCR and ABL genes in Ph1 positive acute leukemias with rearrangements in the BCR gene first intron: Further evidence implicating *Alu* sequences in the chromosome translocation. *Nucleic Acids Res* **17**:7631–7642, 1989.

46. Super HG, Strissel PL, Sobulo OM, Burian D, Reshmi SC, Roe B, Zeleznik-Le NJ, Diaz MO, Rowley JD. Identification of complex genomic breakpoint junctions in the t(9;11) MLL-*AF9* fusion gene in acute leukemia. *Genes Chromosomes Cancer* **20**:185–195, 1997.

47. Hasty P, Rivera-Perez J, Bradley A. The length of homology required for gene targeting in embryonic stem cells. *Mol Cell Biol* **11**:5586–5591, 1991.

48. Burwinkel B, Kilimann MW. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J Mol Biol* **277**:513–517, 1998.

49. Shen M, Deininger P. An *in vivo* assay for measuring the recombination potential between DNA sequences in mammalian cells. *Anal Biochem* **205**:83–89, 1992.

50. Kazazian HH, Jr. Mobile elements and disease. *Curr Opin Genet Dev* **8**:343–350, 1998.

51. Janicic N, Pausova Z, Cole DE, Hendy GN. Insertion of an *Alu* sequence in the Ca(2+)-sensing receptor gene in familial hypocalciuric hypercalcemia and neonatal severe hypoparathyroidism. *Am J Hum Genet* **56**:880–886, 1995.

52. Economou-Pachnis A, Tsichlis PN. Insertion of an *Alu* SINE in the human homologue of the *Mlv1-2* locus. *Nucleic Acids Res* **13**:8379–8387, 1985.

53. Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS. A de novo *Alu* insertion results in neurofibromatosis type 1. *Nature* **353**:864–866, 1991.

54. Rowe SM, Coughlan SJ, McKenna NJ, Garrett E, Kieback DG, Carney DN, Headon DR. Ovarian carcinoma-associated TaqI restriction fragment length polymorphism in intron G of the progesterone receptor gene is due to an *Alu* sequence insertion. *Cancer Res* **55**:2743–2745, 1995.

55. Lester T, McMahon C, VanRegemorter N, Jones A, Genet S. X-linked immunodeficiency caused by insertion of *Alu* repeat sequences. *J Med Gen Suppl* **34**(Suppl 1): S81, 1997.

56. Vidaud D, Vidaud M, Bahnak BR, Siguret V, Gispert Sanchez S, Laurian Y, Meyer D, Goossens M, Lavergne JM. Haemophilia B due to a de novo insertion of a human-specific *Alu* subfamily member within the coding region of the factor IX gene. *Eur J Hum Genet* **1**:30–36, 1993.

57. Abdelhak S, Kalatzis V, Heilig R, Compain S, Samson D, Vincent C, Levi-Acobas F, Cruaud C, Le Merrer M, Mathieu M, Konig R, Vigneron J, Weissenbach J, Petit C, Weil D. Clustering of mutations responsible for branchio-oto-renal (BOR) syndrome in the eyes absent homologous region (eyaHR) of EYA1. *Hum Mol Genet* **6**:2247–2255, 1997.

58. Muratani K, Hada T, Yamamoto Y, Kaneko T, Shigeto Y, Ohue T, Furuyama J, Higashino K. Inactivation of the cholinesterase gene by *Alu* insertion: Possible mechanism for human gene transposition. *Proc Natl Acad Sci USA* **88**:11315–11319, 1991.

59. Halling KC, Honchel R, Lazzaro CR, Bufill JA, Arndt C, Lindor NM. A germline *Alu*1 repeat insertion of the APC gene leading to hereditary desmoid disease in an Amish family. *Am J Hum Genet Suppl* **61**:A67, 1997.

60. Stoppa-Lyonnet D, Carter PE, Meo T, Tosi M. Clusters of intragenic *Alu* repeats predispose the human C1 inhibitor locus to deleterious rearrangements. *Proc Natl Acad Sci* **87**:1551–1555, 1990.

61. Miki Y, Katagiri T, Kasumi F, Yoshimoto T, Nakamura Y.

Mutation analysis in the BRCA2 gene in primary breast cancers. *Nature Genet* **13**:245–247, 1996.

62. Zhang Y-H, Huang B-L, Finlayson G, Deininger PL, McCabe ERB. Alu Sx insertion in a patient with benign glycerol kinase deficiency. *Am J Hum Genet Suppl* **63**:A395, 1998.
63. Chae JJ, Park YB, Kim SH, Hong SS, Song GJ, Han KH, Namkoong Y, Kim HS, Lee CC. Two partial deletion mutations involving the same Alu sequence within intron 8 of the LDL receptor gene in Korean patients with familial hypercholesterolemia. *Hum Genet* **99**:155–163, 1997.
64. Lehrman MA, Goldstein JL, Russel DW, Brown MS. Duplication of seven exons in LDL receptor gene caused by Alu-Alu recombination in a subject with familial hypercholesterolemia. *Cell* **48**:827–835, 1987.
65. Lehrman MA, Schneider WJ, Sudhof TC, Brown MS, Goldstein JL, Russell DW. Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science* **227**:140–146, 1985.
66. Rudiger NS, Heinsvig EM, Hansen FA, Faergeman O, Bolund L, Gregersen N. DNA deletions in the low density lipoprotein (LDL) receptor gene in Danish families with familial hypercholesterolemia. *Clin Genet* **39**:451–462, 1991.
67. Yamakawa K, Takada K, Yanagi H, Tsuchiya S, Kawai K, Nakagawa S, Kajiyama G, Hamaguchi H. Three novel partial deletions of the low-density lipoprotein (LDL) receptor gene in familial hypercholesterolemia. *Hum Genet* **82**:317–321, 1989.
68. Flint J, Rochette J, Craddock CF, Dode C, Vignes B, Horsley SW, Kearney L, Buckle VJ, Ayyub H, Higgs DR. Chromosomal stabilisation by a subtelomeric rearrangement involving two closely related Alu elements. *Hum Mol Genet* **5**:1163–1169, 1996.
69. Ko TM, Tseng LH, Kao CH, Lin YW, Hwa HL, Hsu PM, Li SF, Chuang SM. Molecular characterization and PCR diagnosis of Thailand deletion of alpha-globin gene cluster. *Am J Hematol* **57**:124–130, 1998.
70. Harsteveld KL, Losekoot M, Fodde R, Giordano PC, Bernini LF. The involvement of Alu repeats in recombination events at the alpha-globin gene cluster: Characterization of two alphazero-thalassae mia deletion breakpoints. *Hum Genet* **99**:528–534, 1997.
71. Nicholls RD, Fischel-Ghodsian N, Higgs DR. Recombination at the human alpha-globin gene cluster: Sequence features and topological constraints. *Cell* **49**:369–378, 1987.
72. Ariga T, Carter PE, Davis AE. Recombinations between Alu repeat sequences that result in partial deletions within the C1 inhibitor gene. *Genomics* **8**:607–613, 1990.
73. Poussi B, Hautala T, Heikkilä J, Pajunen L, Kivirikko KI, Myllyla R. Alu-Alu recombination results in a duplication of seven exons in the lysyl hydroxylase gene in a patient with the type VI variant of Ehlers-Danlos syndrome. *Am J Hum Genet* **55**:899–906, 1994.
74. Hu XY, Ray PN, Worton RG. Mechanisms of tandem duplication in the Duchenne muscular dystrophy gene include both homologous and nonhomologous intrachromosomal recombination. *EMBO J* **10**:2471–2477, 1991.
75. Markert ML, Hutton JJ, Wigington DA, States JC, Kaufman RE. Adenosine deaminase (ADA) deficiency due to deletion of the ADA gene promoter and first exon by homologous recombination between two Alu elements. *J Clin Invest* **81**:1323–1327, 1988.
76. Huang LS, Ripp ME, Korman SH, Deckelbaum RJ, Breslow JL. Hypobetalipoproteinemia due to an apolipoprotein B gene exon 21 deletion derived by Alu-Alu recombination. *J Biol Chem* **264**:11394–11400, 1989.
77. Shimada F, Taira M, Suzuki Y, Hashimoto N, Nozaki O, Tatibana M, Ebina Y, Tawata M, Onaya T. Insulin-resistant diabetes associated with partial deletion of insulin-receptor gene. *Lancet* **335**:1179–1181, 1990.
78. Kornreich R, Bishop DF, Desnick RJ. Alpha-galactosidase A gene rearrangements causing Fabry disease. Identification of short direct repeats at breakpoints in an Alu-rich gene. *J Biol Chem* **265**:9319–9326, 1990.
79. Marcus S, Hellgren D, Lambert B, Fallstrom SP, Wahlstrom J. Duplication in the hypoxanthine phosphoribosyl-transferase gene caused by Alu-Alu recombination in a patient with Lesch Nyhan syndrome. *Hum Genet* **90**:477–482, 1993.
80. Li L, Bray PF. Homologous recombination among three intragene Alu sequences causes an inversion-deletion resulting in the hereditary bleeding disorder Glanzmann thrombasthenia. *Am J Hum Genet* **53**:140–149, 1993.
81. Burwinkel B, Kilimann MW. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J Mol Biol* **277**:513–517, 1998.
82. Hori T, Tomatsu S, Nakashima Y, Uchiyama A, Fukuda S, Sukegawa K, Shimozawa N, Suzuki Y, Kondo N, Horiuchi T. Mucopolysaccharidosis type IVA: Common double deletion in the N-acetylgalactosamine-6-sulfatase gene (GALNS). *Genomics* **26**:535–542, 1995.
83. Olds RJ, Lane DA, Chowdhury V, De SV, Leone G, Thein SL. Complete nucleotide sequence of the antithrombin gene: Evidence for homologous recombination causing thrombophilia. *Biochemistry* **32**:4216–4224, 1993.
84. Rouyer F, Simmler MC, Page DC, Weissenbach J. A sex chromosome rearrangement in a human XX male caused by Alu-Alu recombination. *Cell* **51**:417–425, 1987.
85. Myerowitz R, Hogikyan ND. A deletion involving Alu sequences in the beta-hexosaminidase alpha-chain gene of French Canadians with Tay-Sachs disease. *J Biol Chem* **262**:15396–15399, 1987.
86. Botto M, Fong KY, So AK, Barlow R, Routier R, Morley BJ, Walport MJ. Homozygous hereditary C3 deficiency due to a partial gene deletion. *Proc Natl Acad Sci USA* **89**:4957–4961, 1992.
87. Neote K, McInnes B, Mahuran DJ, Gravel RA. Structure and distribution of an Alu-type deletion mutation in Sandhoff disease. *J Clin Invest* **86**:1524–1531, 1990.
88. Strout MP, Marcucci G, Bloomfield CD, Caligiuri MA. The partial tandem duplication of ALL1 (MLL) is consistently generated by Alu-mediated homologous recombination in acute myeloid leukemia. *Proc Natl Acad Sci USA* **95**:2390–2395, 1998.
89. So CW, Ma ZG, Price CM, Dong S, Chen SJ, Gu LJ, So CK, Wiedemann LM, Chan LC. MLL self fusion mediated by Alu repeat homologous recombination and prognosis of AML-M4/M5 subtypes. *Cancer Res* **57**:117–122, 1997.
90. Schichman SA, Caligiuri MA, Strout MP, Carter SL, Gu Y,

Canaani E, Bloomfield CD, Croce CM. ALL-1 tandem duplication in acute myeloid leukemia with a normal karyotype involves homologous recombination between Alu elements. *Cancer Res* **54**:4277-4280, 1994.

91. Swensen J, Hoffman M, Skolnick MH, Neuhausen SL. Identification of a 14 kb deletion involving the promoter region of BRCA1 in a breast cancer family. *Hum Mol Genet* **6**:1513-1517, 1997.

92. Puget N, Sinilnikova O, Stoppa-Lyonnet D, Ayyadevera R, Pages S, Lynch H, Goldgar D, Lenoir GM, Mazoyer S. An Alu-mediated 6-kb duplication in the BRCA1 gene: A new founder mutation? *Am J Hum Genet* **64**:300-302, 1999.

93. Mauillon JL, Michel P, Limacher JM, Latouche JB, Dechelotte P, Charbonnier F, Martin C, Moreau V, Metayer J, Paillot B, Frebourg T. Identification of novel germline hMLH1 mutations including a 22 kb Alu-mediated deletion in patients with familial colorectal cancer. *Cancer Res* **56**: 5728-5733, 1996.

94. Onno M, Nakamura T, Hilleva J, Hill M. Rearrangement of the human tre oncogene by homologous recombination between Alu repeats of nucleotide sequences from two different chromosomes. *Oncogene* **7**:2519-2523, 1992.

95. Rothberg PG, Ponnuru S, Baker D, Bradley JF, Freeman AI, Cibis GW, Harris DJ, Heruth DP. A deletion polymorphism due to Alu-Alu recombination in intron 2 of the retinoblastoma gene: Association with human gliomas. *Mol Carcinog* **19**:69-73, 1997.

96. Zucman-Rossi J, Batzer MA, Stoneking M, Delattre O, Thomas G. Interethnic polymorphism of EWS intron 6: Genome plasticity mediated by Alu retrotransposition and recombination. *Hum Genet* **99**:357-363, 1997.

Recently integrated human Alu repeats: finding needles in the haystack

Astrid M. Roy¹, Marion L. Carroll², David H. Kass³, Son V. Nguyen², Abdel-Halim Salem^{2,*}, Mark A. Batzer² & Prescott L. Deininger^{1,4,**}

¹*Tulane Cancer Center, SL-66, Department of Environmental Health Sciences, Tulane University - Medical Center, 1430 Tulane Ave., SL-66, New Orleans, LA 70112, USA;* ²*Departments of Pathology, Biometry and Genetics, Biochemistry and Molecular Biology, Stanley S. Scott Cancer Center, Neuroscience Center of Excellence, Louisiana State University Health Sciences Center, 1901 Perdido Street, New Orleans, LA 70112, USA;* ³*Department of Biology, 316 Mark Jefferson, Eastern Michigan University, Ypsilanti, MI 48197, USA;* ^{*}*Present address: Department of Anatomy, Faculty of Medicine, Suez Canal University, Ismailia, Egypt;* ⁴*Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, 1516 Jefferson Highway, New Orleans, 70121;* ^{**}*Author and address for correspondence: Tulane Cancer Center, Tulane University – Medical Center, 1430 Tulane Ave., SL-66, New Orleans, LA 70112, USA (Phone: (504) 988-6385; Fax: (504) 588-5516; E-mail: pdeinin@tcs.tulane.edu)*

Accepted 25 January 2000

Key words: Alu insertion polymorphisms, anchored-PCR, comparative genomics, computational biology

Abstract

Alu elements undergo amplification through retroposition and integration into new locations throughout primate genomes. Over 500,000 Alu elements reside in the human genome, making the identification of newly inserted Alu repeats the genomic equivalent of finding needles in the haystack. Here, we present two complementary methods for rapid detection of newly integrated Alu elements. In the first approach we employ computational biology to mine the human genomic DNA sequence databases in order to identify recently integrated Alu elements. The second method is based on an anchor-PCR technique which we term Allele-Specific Alu PCR (ASAP). In this approach, Alu elements are selectively amplified from anchored DNA generating a display or 'fingerprint' of recently integrated Alu elements. Alu insertion polymorphisms are then detected by comparison of the DNA fingerprints generated from different samples. Here, we explore the utility of these methods by applying them to the identification of members of the smallest previously identified subfamily of Alu repeats in the human genome termed Ya8. This subfamily of Alu repeats is composed of about 50 elements within the human genome. Approximately 50% of the Ya8 Alu family members have inserted in the human genome so recently that they are polymorphic, making them useful markers for the study of human evolution.

Introduction

Alu repeats are the most successful class of mobile elements in the human genome. Alu elements spread through the genome via an RNA mediated amplification mechanism termed retroposition and reviewed in Deininger and Batzer, 1993. There are over 500,000 Alu elements in the human genome, which have clearly played a major role in sculpting and/or damaging the genome. Alu elements have contributed to genetic disease, both by the disruption of genes through the insertion of newly retroposed ele-

ments and by recombination between Alu elements (reviewed in Deininger & Batzer, 1999). Previous estimates indicate that retroposition of Alu elements contributes to approximately 0.1% of human genetic diseases and recombination between Alu repeats contributes to another 0.3% of genetic diseases (Deininger & Batzer, 1999). Therefore, the spread of the Alu family of mobile elements has generated a significant amount of human genomic variation as well as diseases through recombination-based fluidity as well as insertional mutagenesis.

Alu repeats are distributed rather haphazardly throughout the human genome. Alu elements began expanding in the ancestral primate genomes about 65 mya (Shen, Batzer & Deninger, 1991) reaching a peak amplification between 35 and 60 mya. Presently, Alu elements amplify at a rate that is 100 fold lower than their peak rate, with an estimate of one new Alu insert in every 100–200 births (Deininger & Batzer, 1993, 1995). Evolutionary studies have demonstrated that the majority of evolutionarily recent Alu inserts have specific diagnostic sequence mutations (Deininger & Batzer, 1993, 1995). These mutations have accumulated in Alu elements throughout primate evolution resulting in a hierarchical subfamily structure, or lineage, of Alu repeats. The mutations facilitate the classification of Alu elements into different subfamilies, or clades, of related elements that share common diagnostic mutations (reviewed in Batzer, Schmid & Deninger, 1993; Batzer & Deininger, 1991; Batzer et al., 1996a). Almost all of the recently integrated Alu elements within the human genome belong to one of four closely related subfamilies: Y, Ya5, Ya8, and Yb8, with the majority being Ya5 and Yb8 subfamily members. Collectively, these subfamilies of Alu elements comprise less than 10% of the Alu elements present within the human genome with the Ya5/8 and Yb8 subfamilies collectively accounting for less than half of a percent of all Alu elements. These evolutionarily recent Alu insertions are useful for human population studies, since there appears to be no specific mechanism to remove newly inserted Alu repeats, and the Alu elements are identical by descent with a known ancestral state (Batzer et al., 1991, 1994a, 1996a; Stoneking et al., 1997; Perna et al., 1992).

Previously, it has been technically impossible to determine the full impact of mobile elements on the human genome. The identification of newly inserted Alu elements has been very difficult due to the complexity of detecting one new Alu insertion in a cell that already has 500,000 pre-existing Alu elements. We have previously utilized laborious library screening and sequencing strategies to isolate relatively small numbers of Alu insertion polymorphisms (Arcot et al., 1995a, b, c; Batzer & Deininger 1991a; Batzer et al., 1990, 1991b; 1995), as well as investigating rare 300 bp restriction fragment length polymorphisms (Kass et al., 1994). This makes these studies the genomic equivalent of the search for needles in the haystack. In this paper, we discuss two alternative methods that overcome the inherent difficulties in these experiments, making these studies manage-

able. First, the availability of large quantities of human genomic DNA sequence provided by the Human Genome Project facilitates genomic database mining for recently integrated Alu elements. This approach should prove useful in determining the chromosome-specific and genome wide dispersal patterns of mobile elements, as well as for the identification of polymorphic mobile element fossils to apply to the study of human population genetics and primate comparative genomics. Secondly, we have developed a PCR-based method that we term Allele-Specific Alu PCR (ASAP). This technique allows us to take advantage of the subfamily-specific diagnostic mutations within Alu mobile elements to isolate and display recently integrated Alu repeats from different DNA samples, allowing for direct comparisons of the Alu content of different genomes or different cells from an individual.

Materials and methods

Cell lines and DNA samples

The cell lines used to isolate human DNA samples were as follows: human (*Homo sapiens*), HeLa (ATCC CCL2); chimpanzee (*Pan troglodytes*), Wes (ATCC CRL1609), gorilla (*Gorilla gorilla*), Ggo-1 (primary gorilla fibroblasts) provided by Dr. Stephen J. O'Brien, National Cancer Institute, Frederick, MD, USA. Cell lines were maintained as directed by the source and DNA isolations were performed using Wizard genomic DNA purification (Promega). Human DNA samples from the European, African American and Greenland native population groups were isolated from peripheral blood lymphocytes (Ausubel et al., 1996) that were available from previous studies (Stoneking et al., 1997). Egyptian samples were collected from throughout the Nile river valley region and DNA from peripheral lymphocytes was prepared using Wizard genomic DNA purification kits (Promega). Human DNA used for ASAP was isolated from peripheral lymphocytes utilizing the super-quick gene method (Analytical Genetic Testing Center).

Computational analyses

A schematic overview summarizing the computational analyses of recently integrated Alu elements is shown in Figure 1. Initial screening of the GenBank non-redundant and high throughput genomic sequence (HTGS) databases was performed using the basic local

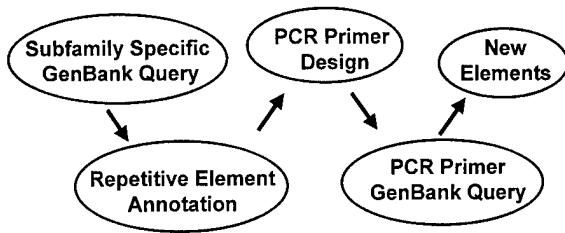


Figure 1. Computational analysis of repetitive elements. The flow chart shows the computational tools utilized for the identification and analysis of recently integrated Ya8 Alu family members. The process begins with BLAST searches of the non-redundant and high-throughput genomic sequence databases. Subsequently sequences (about 1000 nucleotides) adjacent to the matches with 100% identity to the query sequence are annotated using the Repeat-Masker2 or Censor server. Following sequence annotation, oligonucleotide primers complementary to the unique DNA sequences adjacent to each element are designed using the Primer3 web server. The oligonucleotides designed using Primer3 are then subjected to a second BLAST search to determine if they reside in other repetitive elements, and subsequently they are used for PCR based analyses of individual mobilé elements.

alignment search tool (BLAST) (Altschul et al., 1990) available from the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). The database was searched for exact complements to the oligonucleotide 5'-ACTAAAACACTACAAAAAAATAG-3' that is an exact match to a portion of the Alu Ya8 subfamily consensus sequence containing unique diagnostic mutations. Sequences that were exact complements to the oligonucleotide were then subjected to more detailed annotation. A region composed of 1000 bases of flanking DNA sequence directly adjacent to the sequences identified from the databases that matched the initial GenBank BLAST query were subjected to annotation using either Repeat-Masker2 from the University of Washington Genome Center server (<http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>) or Censor from the Genetic Information Research Institute (http://www.girinst.org/Censor_Server-Data_Entry_Form_s.html) (Jurka et al., 1996). These programs annotate the repeat sequence content of DNA sequences from humans and rodents.

Primer design and PCR amplification

PCR primers were designed from flanking unique DNA sequences adjacent to individual Ya8 Alu elements using the Primer3 software (Whitehead Institute for Biomedical Research, Cambridge, MA, USA) (http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). The resultant PCR primers were screened against the GenBank non-redundant data-

base for the presence of repetitive elements using the BLAST program, and primers that resided within known repetitive elements were discarded and new primers were designed. PCR amplification was carried out in 25 μ l reactions using 50–100 ng of target DNA, 40 pM of each oligonucleotide primer, 200 μ M dNTPs in 50 mM KCl, 1.5 mM MgCl₂, 10 mM Tris-HCl pH 8.4 and Taq[®] DNA polymerase (1.25 U) as recommended by the supplier (Life Technologies). Each sample was subjected to the following amplification cycle: an initial denaturation of 2:30 min at 94°C, 1 min of denaturation at 94°C, 1 min at the annealing temperature, 1 min of extension at 72°C, repeated for 32 cycles, followed by a final extension at 72°C for 10 min. Twenty microliters of each sample was fractionated on a 2% agarose gel with 0.25 μ g/ml ethidium bromide. PCR products were directly visualized using UV fluorescence. The sequences of the oligonucleotide primers, annealing temperatures, PCR product sizes and chromosomal locations are shown in Table 1. Phylogenetic analysis of all the Alu elements listed in Table 1 was determined by PCR amplification of human and non-human primate DNA samples. The human genomic diversity associated with each element was determined by the amplification of 20 individuals from each of four populations (African-American, Greenland Native, European and Egyptian) (160 total chromosomes). The chromosomal location of Alu repeats identified from clones that had not been previously mapped was determined by PCR amplification of National Institute of General Medical Sciences (NIGMS) human/rodent somatic cell hybrid mapping panel 2 (Coriell Institute for Medical Research, Camden, NJ).

Allele-Specific Alu PCR (ASAP)

We used a modification of the IRE-Bubble PCR method (Munroe et al., 1994), utilizing the same amplification (anchor) primer, but altering the annealed anchor/linker primers. The annealed linkers formed a Y instead of a bubble to avoid end-to-end ligation. Also, instead of blunt-end digestion, genomic DNA was digested with *Mse*I; that cuts 5'-TAA-3' and does not cut in the Alu consensus. Otherwise the genomic-anchor ligations were prepared according to (Munroe et al., 1994). The annealed linker primers are: MSET: 5'-TAGAAGGAGAGG-ACGCTGTCTGTCGAAGG-3' and MSEB: 5'-GAGCGAATTCGTCAACATAGCATTCTGTCCTCTCC TTC-3'. The amplification (linker) primer is: LNP:

Table 1. Ya8 accession numbers, primers, location, and product sizes

Name	Accession #	5' Primer sequence (5'-3')	3' Primer sequence (5'-3')	A. T. ¹	Chromosomal location ²	Product size ³	
						Filled	Empty
Ya8NBC1	AC006959	CCTGCTGACATTAGAAATGACTT	ATATACAAGTCATCAGATGGGGACAC	60°C	5	504	293
Ya8NBC2	AC006556	GCCTGTGTACCTCCTTTAAATATCTTG	CTCAAAAACACTGGAGCAGGGATCAA	50°C	21	503	242
Ya8NBC3	AC006989	GGTGGTCATCCATATACTATCTCATAGG	AGAGTTCGGAAAAGTGTGACAGGAT	55°C	Y/X ⁴	498	178
Ya8NBC4	AL049871	CATTCCACCCCTGTCAAGCATT	GCTTTGGAAACTAGGCAGGCTTAC	60°C	14	536	204
Ya8NBC6	AC004066	ACTTAGCTTGTAGTATTTCTGAACATTC	CTAAAATGGAGGTACCGATACTTTATA	60°C	4	470	132
Ya8NBC8	AL034422	GGATTCACAAACCTAAATGAAAGAGTAA	CCGTCTCAAAACAAAGAGCAAATA	60°C	20	501	155
Ya8NBC10	AC004893	GGATTACTTTGATGAAAATATCTTAGTAGG	AACTGGATGTGACTTTGAAGACAC	60°C	7	757	371
Ya8NBC11	AC007688	GAGTGCCCTTATTATGTGTTAGGTACITITGCT	ACTCTCACTAGATTATAAGCCCCATAAGGA	60°C	12	419	105
Ya8NBC12	AL022302	CATCTTAAAGAACATTAAGAAAAGTACACAG	CTGGCCCACTTACATGATATTTCATACTAG	60°C	22	530	211
Ya8NBC13	AL008722	CCATTTCATTAAGAAGGCTTCAACC	AAAGTAATGTAAGAAAGTATTGGAGAAAGAGAT	60°C	22	402	77
Ya8NBC14	AF094481	GAATCTCTATCTGTACACTAGCCACT	GGCAAACAAAGTCTGTAGTAATCTAAAGGAG	60°C	3	500	189
Ya8NBC15	AF179296	CTCTACAGTACAGATGAGAAAAGTACAGACA	CGCCCTTGCTAGATTTCCTTCTAATCG	60°C	8	620	299
Ya8NBC17	AC005205	CTAGTTCCCACATACCGAAAACAC	CCTGTCTCGTTCAAGTCAGTCCTCTTGT	58°C	19	501	155
Ya8NBC60	AC006553	CAGTCCCATAGCAGTCATGGTAAATAAG	AAGTCTATACCGGTTACCTCTTCTT	58°C	4	456	149

¹ Amplification of each locus required 2:30 min @ 94°C initial denaturing, and 32 cycles for 1 min 94°C, 1 min Annealing Temperature (A.T.) and 1 min elongation at 72°C, with a final extension time of 10 min at 72°C.

² Chromosomal location determined from Accession information or by PCR analysis of monochromosomal hybrid cell lines.

³ Empty product sizes calculated by removing the Alu element and one direct repeat from the filled sites that were identified.

⁴ Ya8NBC3 is located in the pseudoautosomal region of the X and Y chromosome.

5'GAATTCGTCAACATAGCATTCT-3'. We placed an *Eco*RI site at the 5' end of the primer for the option of cloning PCR products into cloning sites of common vectors. No bands are observed on a gel when this primer is used alone with the anchored template at an annealing temperature of 55°C.

Unless otherwise noted, PCR conditions (for all ASAP reactions) were performed in 20 µl using a Perkin-Elmer 9600 thermal cycler with the following conditions: 1 × Promega buffer, 1.5 mM MgCl₂, 200 µM dNTPs, 0.25 µM primers, 1.5 U Taq polymerase (Promega) at 94°C – 2 min, 94°C – 20 s, 62°C – 20 s, 72°C – 1 min, 10 s, for 5 cycles; 94°C – 20 s, 55°C – 20 s, 72°C – 1 min, 10 s, for 25 cycles; 72°C – 3 min. Nested Alu primers were used that move along the Alu in an upstream direction as follows: ASII (Ya5-specific): 5'-CTGGAGTCAGTGGCGG-3'; HS18R (Ya8-specific): 5'-CTCAGCCTCCCAAGTAGCTA-3'; HS16R (Ya8-specific): 5'-CGCCCGGCTATTTTGTAG-3'.

The ASII primer has Ya5 diagnostic nucleotides (present in both Ya5 and Ya8 subfamilies). In the first round of PCR, stock genomic DNA (2.4 ng anchored DNA) was used as the template. For subsequent rounds of amplification, PCR products were purified through microcon-30 (Amicon) columns using two centrifuge spins following the addition of 400 µl of water. For the second round of amplification, 1 µl of microcon-purified first round PCR reaction was used as the template, and for the third round 1 µl of microcon-purified second round PCR products was used. For display analysis (see below) the PCR products were 'equalized' in volume following microcon purification.

Display of anchor-Alu PCR products

Third round PCR was performed utilizing a 5' end-labeled primer incorporating [γ -³²P] ATP (Amersham) with T4 polynucleotide kinase (New England BioLabs). PCR conditions were as above with the exception of using 0.188 µM of each Ya8 and LNP cold primers and 0.075 µM of end-labeled Ya8 primer. Anchor-PCR and end-labeled molecular weight markers (φX174 DNA digested with *Hinf*I; Promega) were separated by electrophoresis on denaturing 5% long ranger (AT Biochem) gels, and examined by autoradiography following exposure to Amersham Hyperfilm at room temperature. DNA samples from different ethnic groups were utilized in the display to identify

variants that resulted from recent Alu insertion events (polymorphism).

Verification of PCR generated DNA fragments as Ya8 products

Gels were aligned to autoradiographs by either small cuts in various parts of the gel, or placement of low-level radioactive dye on the gel prior to re-exposure. Bands were then sliced out of the gels, placed in 200 µl of water and eluted by heating at 65°C for 15 min. Samples were re-amplified with third round PCR primers, cloned and sequenced as described above. Following verification these bands were amplified by the third round primer pair, new nested oligonucleotides based on the flanking unique sequences were designed to move, by PCR, downstream through the Alu element to the opposite flank. Annealing temperatures were adjusted to reflect the Tm of the oligonucleotide primers. Generally two or three rounds of PCR were utilized to obtain the 3' flanking sequences of the Alu. These PCR products were also cloned and sequenced in the same manner.

Results

We present two complementary approaches that facilitate rapid detection of newly inserted Alu elements from the human genome. First, computational analyses of human genomic DNA sequences from the GenBank database are used in the identification of recently integrated Alu elements. Second, allele-specific PCR amplification is used for the selective enrichment of young Alu elements. To compare and contrast these two approaches, we present the data obtained when these methods are applied to the identification of members of the Ya8 Alu subfamily, the smallest previously reported subfamily of Alu repeats in the human genome.

Copy number and sequence diversity

In order to estimate the copy number of Ya8 Alu family members, we determined the number of exact matches to our subfamily specific oligonucleotide query sequence as a proportion of the human genome that had been sequenced in the non-redundant database. We obtained 27 matches to the subfamily specific query sequence from the non-redundant database. Upon further sequence annotation using the RepeatMasker2 web site, five matched the Ya8 Alus

previously sequenced in our laboratories (Batzer et al., 1990; Batzer & Deininger, 1991; Batzer et al., 1995). Eight of the elements identified in the search were classified as Alu Sx subfamily members, and two matched the TPA 25 Ya8 Alu family member. A total of 13 independent Ya8 Alu elements were identified from the search of the non-redundant database that were not sequenced as part of a project to specifically identify recently integrated Alu elements. The non-redundant database contained 45.3% human DNA sequences for a total of 590,140,703 bases of human sequence on the date of the search. The estimated size of the Ya8 subfamily is $(3 \times 10^9 \text{ bp}/590,140,703 \text{ bp}) \times 13 \text{ unique Ya8 matches} = 66 \text{ Ya8 subfamily members}$. This estimate compares favorably with that of 50 previously reported based upon library screening, restriction digestion or Southern blotting (Batzer et al., 1995). An additional six matches to the Ya8 subfamily query sequence were identified in the HTGS. One of these elements was an Alu Sq subfamily member, while a second element was a duplicate copy of Ya8NBC60. PCR analyses of two elements identified in the high throughput database, Ya8NBC7 and Ya8NBC16 (GenBank accession numbers AL109937 and AC008944), were inconclusive and these elements were eliminated from further analysis. These two elements were identified from low pass first sequence runs in the HTGS database. It is not surprising that the PCR analyses failed, since the DNA sequences are of presumably lower quality than finished DNA sequences contained in the non-redundant database. However, two additional Ya8 Alu repeats (Ya8NBC8 and Ya8NBC15) were identified in the HTGS database and subjected to further analysis.

A comparison of the nucleotide sequences of all of the Ya8 Alu family members is shown in Figure 2. In order to determine the time of origin for the Ya8 subfamily we divided the nucleotide substitutions within the elements into those that have occurred in CpG dinucleotides and those that have occurred in non-CpG positions. The distinction between types of mutations is made because the CpG dinucleotides mutate at a rate that is about 10 times faster than non-CpG positions (Labuda & Striker, 1989; Batzer et al., 1990) as a result of the deamination of 5-methylcytosine (Bird, 1980). A total of 14 non-CpG mutations and 8 CpG mutations occurred within the 14 Alu Ya8 subfamily members reported. Using a neutral rate of evolution for primate intervening DNA sequences of 0.15% per million years (Miyamoto, Slightom & Goodman, 1987) and the non-CpG mutation rate of 0.413%

AluYa8 Con	GGCCGGCCCGGTGGCTCACGCCCTGTAACTCCAGACACTTTGGGAGGCCGAGCGGGCGGG	59
AluYa8NBC1	
AluYa8NBC2	
AluYa8NBC3	
AluYa8NBC4	
AluYa8NBC6	
AluYa8NBC8	
AluYa8NBC10	
AluYa8NBC11	
AluYa8NBC12	
AluYa8NBC13	
AluYa8NBC14	
AluYa8NBC15	
AluYa8NBC17	
AluYa8NBC60	
AluYa8 Con	ATCACGAGGTCAAGGAGATCGAGACCATCCGGCTAAAAACCGTGAACCCCGCTCTACT	118
AluYa8NBC1	
AluYa8NBC2	
AluYa8NBC3	
AluYa8NBC4	
AluYa8NBC6	
AluYa8NBC8	
AluYa8NBC10	
AluYa8NBC11	
AluYa8NBC12	
AluYa8NBC13	
AluYa8NBC14	
AluYa8NBC15	
AluYa8NBC17	
AluYa8NBC60	
AluYa8 Con	AAAACTAACAAAAATAGCCGGGGTAGTGGCGGGGCCGTAGTCTAGCTACTTTGGG	177
AluYa8NBC1	
AluYa8NBC2	
AluYa8NBC3	
AluYa8NBC4	
AluYa8NBC6	
AluYa8NBC8	
AluYa8NBC10	
AluYa8NBC11	
AluYa8NBC12	
AluYa8NBC13	
AluYa8NBC14	
AluYa8NBC15	
AluYa8NBC17	
AluYa8NBC60	
AluYa8 Con	GCGTGAGGCAGGAAATGGCTAAACCGGGAGGGAGCTTCCAGTGAGCCGAGATCCC	237
AluYa8NBC1	
AluYa8NBC2	
AluYa8NBC3	
AluYa8NBC4	
AluYa8NBC6	
AluYa8NBC8	
AluYa8NBC10	
AluYa8NBC11	
AluYa8NBC12	
AluYa8NBC13	
AluYa8NBC14	
AluYa8NBC15	
AluYa8NBC17	
AluYa8NBC60	
AluYa8 Con	GCCACTGGACTCCAGCTGGGGACAGAGCGAGACTCCGTCAAAAAAA	290
AluYa8NBC1	
AluYa8NBC2	
AluYa8NBC3	
AluYa8NBC4	
AluYa8NBC6	
AluYa8NBC8	
AluYa8NBC10	
AluYa8NBC11	
AluYa8NBC12	
AluYa8NBC13	
AluYa8NBC14	
AluYa8NBC15	
AluYa8NBC17	
AluYa8NBC60	

Figure 2. Multiple alignment of Ya8 subfamily members. The Ya8 subfamily consensus (con) is derived from the most common nucleotide found at each position within the subfamily members. Nucleotide substitutions at each position are indicated with the appropriate nucleotide. Deletions are marked by '-'.

(14/3388 using only non-CpG bases) within the 14 Ya8 Alu elements yields an estimated age of 2.75 million years old for the Ya8 subfamily members. This estimate of age is somewhat higher than the 660,000 years previously reported (Batzer et al., 1995). However, the previous study of Ya8 Alu family members involved only four elements making the calculated age more subject to random statistical fluctuation. This estimate is also consistent with the expansion of a family of mobile elements that began around the time humans

Ya8NBC1	<u>AAGAGGGGGAGAG</u>	[Alu]	A ₁₈	AAGAGGGGGAGAG
Ya8NBC2	<u>GGG</u>	[Alu]	A ₁₈ ,CA ₂	TGGA
Ya8NBC3	<u>GAAGAAGTTT</u> TG	[Alu]	A ₁₈ ,CA ₂	GAAGAAGTTTGC
Ya8NBC4	<u>CGACAA</u> TTT	[Alu]	A ₁₈ ,CA ₁ ,CA ₁₀	CGACAAATTT
Ya8NBC6	<u>AAATTAA</u> AAATT	[Alu]	A ₁₈	AAATTAAAAATTT
Ya8NBC8	<u>AGAAAAA</u> ATAGCCATA	[Alu]	A ₁₈ ,CA ₁ ,CA ₂	AGAAAAAATAGCCATA
Ya8NBC10	<u>AAA</u> AAATAAAATA	[Alu]	A ₁₈	AAAATAAAATA
Ya8NBC11	<u>AACGAATGAGACTG</u>	[Alu]	A ₂₀	AACGAATGAGACTG
Ya8NBC12	<u>AAA</u> GTCTTGCA	[Alu]	A ₂₇	AAA GTCTTGCA
Ya8NBC13	<u>AGAAAGGCTCACCA</u> G	[Alu]	A ₃₀	AGAAAGGCTCACCA
Ya8NBC14	<u>ATCC</u>	[Alu]	A ₂₆	ATCC
Ya8NBC15	<u>AGAAC</u> CCACAGAA	[Alu]	A ₂₇	AGAACCCACAGAA
Ya8NBC17	<u>AAAGGA</u> ATCTC	[Alu]	A ₁₇	AAAGGAATCTC
Ya8NBC60	<u>GGTAAATAAGCTT</u> CTT	[Alu]	A ₂₅	GGTAAATAAGCTTCTT

Figure 3. Nucleotide sequences flanking Ya8 subfamily members. Nucleotide sequences flanking the Ya8 Alu family members are shown. Nucleotides encompassed in the direct repeats are underlined. The length of the oligo-dA rich tail is denoted by an (A) and a subscript indicating the number of adenine residues.

and African apes diverged, which is thought to have occurred 4–6 million years ago (Miyamoto, Slightom & Goodman, 1987).

Inspection of the nucleotide sequences flanking each Ya8 Alu family member shows that all of the elements were flanked by short perfect direct repeats (Figure 3). The direct repeats ranged in size from 3–17 nucleotides. These direct repeats are fairly typical of recently integrated Alu family members. Two of the Alu Ya8 Alu family members contained 5' truncations (Ya8NBC2 and Ya8NBC11). Since Ya8NBC2 and Ya8NBC11 are both flanked by perfect direct repeats the truncations in these elements probably occurred as a result of incomplete reverse transcription or improper integration into the genome rather than by post-integration instability. All of the Ya8 Alu family members had oligo-dA rich tails that ranged in length from a minimum of four nucleotides to over 40 bases in length. It is also interesting to note that the 3' oligo-dA rich tails of several of the elements (Ya8NBC2, Ya8NBC3, Ya8NBC4, and Ya8NBC8) have accumulated random mutations beginning the process of the formation of simple sequence repeats of varied sequence complexity. The oligo-dA rich tails and middle A rich regions of Alu elements have previously been shown to serve as nuclei for the genesis of simple sequence repeats (Arcot et al., 1995b).

Phylogenetic distribution, and chromosomal location

The phylogenetic distribution of each Ya8 Alu element was determined by amplifying genomic DNA from two non-human primates (common chimpanzee and gorilla). All of the Ya8 Alu family members were absent from the genomes of non-human primates. This suggests that the majority of these elements dispersed within the human genome sometime after the human and African ape divergence. The chromosomal loca-

tion of each Ya8 Alu element was taken directly from the GenBank database entry or determined by PCR amplification of human/rodent monochromosomal hybrid cell line DNA samples (Table 1).

Human genomic diversity

In order to determine the human genomic variation associated with each of the Ya8 Alu family members we subjected a panel of human DNA samples to PCR amplification (Table 2). The panel was composed of 20 individuals of European origin, African Americans, Greenland Natives and Egyptians for a total of 80 individuals (160 chromosomes). Using this approach four of the 14 (Ya8NBC8, Ya8NBC10, Ya8NBC14 and Ya8NBC15) Alu Ya8 subfamily members were monomorphic for the presence of the Alu element suggesting that these elements integrated in the genome prior to the radiation of modern humans from Africa. Three of the elements (Ya8NBC2, Ya8NBC13 and Ya8NBC17) appeared heterozygous in all of the individuals that were analyzed, suggesting that they had integrated into previously undefined repetitive elements within the human genome as previously described (Batzer et al., 1991). However, the remaining seven elements were polymorphic for the presence of an Alu repeat within the genomes of the test panel individuals (Table 2). The unbiased heterozygosity values (corrected for small sample sizes) for these polymorphic Alu insertions were variable, and approached the theoretical maximum in several cases. This is quite interesting since the maximum uncorrected heterozygosity for these biallelic elements is 50% and suggests that these Alu insertion polymorphisms will make excellent markers for the study of human population genetics. In addition, 50% of the randomly identified Ya8 Alu family members are polymorphic. These results suggest that the Ya8 subfamily is younger than either the Ya5 (from which Ya8 was derived) or Yb8 Alu subfamilies, since only 25% of the members of these Alu subfamilies are polymorphic in the human genome (Batzer et al., 1995).

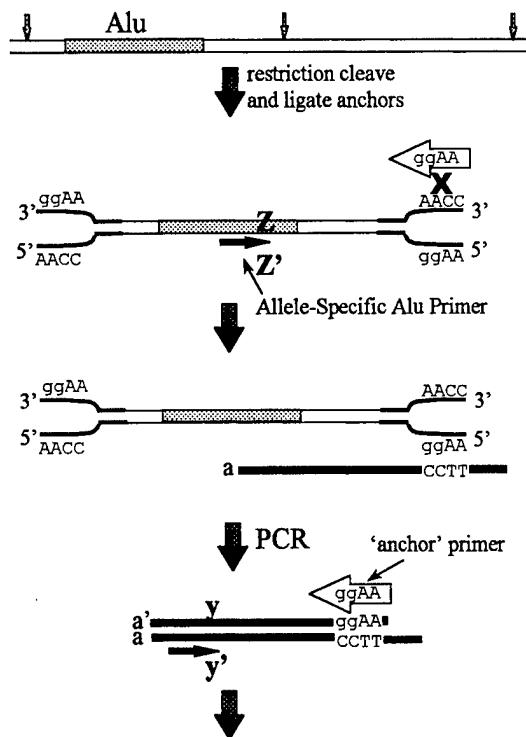
Allele-Specific Alu PCR (ASAP)

Although database screening is extremely efficient for identifying recent Alu elements, it will not allow identification of new elements from genomes not included in the sequencing efforts. Our primary objective with the ASAP technique is to rapidly identify newly inserted Alu elements from a background of 500,000 older Alus. To accomplish this feat, we utilized a

Table 2. Alu Ya8 associated human genomic diversity

Elements	African American			Greenland natives			European			Egyptian		
	Genotypes		<i>f</i> Alu	Het	Genotypes		<i>f</i> Alu	Het	Genotypes		<i>f</i> Alu	Het
	+/+	+/—	—/—	+/—	+/—	—/—	+/+	+/—	—/—	+/+	+/—	—/—
Ya8NBC1	10	2	7	0.58	0.50	5	0	9	0.36	0.35	10	5
Ya8NBC3	0	12	7	0.32	0.44	0	6	14	0.15	0.44	0	12
Ya8NBC4	1	4	13	0.17	0.29	6	0	7	0.46	0.51	8	5
Ya8NBC6	8	2	6	0.56	0.51	11	0	3	0.85	0.00	16	0
Ya8NBC11	13	2	0	0.93	0.13	12	0	0	1.00	0.09	10	1
Ya8NBC12	17	0	0	1.00	0.00	19	0	0	1.00	0.05	18	1
Ya8NBC60	6	9	3	0.58	0.50	6	7	5	0.53	0.51	5	9

¹This is the unbiased heterozygosity.²Average heterozygosity is the average of the population heterozygosity.



Nested Allele-Specific Alu PCR

Figure 4. The Allele-Specific Alu PCR (ASAP) anchor strategy. Schematic diagram of the technique for the isolation of a designated subset of Alu repeats based on a modification of the IRE-bubble PCR technique (Munroe et al., 1994). The shaded rectangle represents an Alu sequence in genomic DNA. The *MseI* (or an alternative restriction enzyme) cleaves in unique sequences flanking the Alu repeat (small arrows). The anchors with the complementary *MseI* site are ligated. The anchors are designed so that the two oligonucleotide strands base-pair only at the *MseI* site end, but not at the other end (represented here schematically with four arbitrary bases). PCR is initiated using an allele-specific Alu primer (*Z'*). The anchor primer will not be able to base pair preventing anchor-to-anchor amplification. Only those fragments (*a*) generated by the Alu primer are available for amplification by the anchor primer. The amplified product (*a* and *a'*) provides a template for nested PCR (primer *y'*) to further decrease the background.

modification of the IRE-bubble PCR technique (Munroe et al., 1994). The procedure utilizes an anchored PCR strategy (Figure 4) in which genomic DNA is cleaved with an enzyme that does not cleave within the Alu repeat. The modified anchor is then ligated to the fragment ends. This anchor will only allow PCR amplification if a primer first primes within the fragment and replicates across the linker eliminating any problems with amplification from anchor to anchor. We take advantage of the base changes that identify the younger Alu subfamily members (Batzer et al., 1996b; Batzer & Deininger, 1991). In addition, this allows

the selective enrichment for a smaller fraction of the Alu elements from the genome, as there are only 1000 Ya5 and 1000 Yb8 Alu repeats and approximately 50 Ya8 Alu family members in the human genome (Batzer et al., 1995). We gain the specificity for the recent inserts by using a PCR primer that matches the particular Alu subfamily with the diagnostic positions at its 3' end. Each amplification will extend from a specific Alu subfamily member through its upstream flanking sequences to the randomly located flanking restriction site. The numerous older Alu repeats have accumulated many mutations and may compete for the PCR primers with the Ya5/8 elements. Therefore, although the first amplification provides a great deal of subfamily specificity, we then carry out a 'nested' reaction using a second allele-specific primer to improve the specificity, followed by a third round with another allele-specific primer. In theory, we can utilize primers for each of the 5–8 diagnostic mutations in a subfamily.

In the example presented in this paper, we focused our attention on the identification and display of the lower copy number Alu Ya8 subfamily. Also, to better display the results, we used nested primers in the upstream direction of Ya8 to avoid amplification problems through the A-rich tail. Using the primers described in the Materials and methods section, by the third round of PCR, we were able to visualize discrete DNA fragments on an agarose gel (data not shown). The size range of these fragments appeared to be between 150 bp and 800 bp. To enhance this display, we chose an alternative method of electrophoretic separation and end-labeled the nested primer to further minimize background (see below). To verify these were Ya8 repeats, we directly cloned the third round PCR products and sequenced them. Partial or complete sequences of these products, using vector primers in both directions, demonstrated all 12 clones to be amplified by the Alu-anchor primer pair, although in one case the unique linker sequence was imprecise. All these elements contained the Ya5/8 diagnostic nucleotides (There were no further upstream diagnostics to declare these as Ya8 elements.).

For eight of the 12 isolated clones, there were between 12 and 18 unique nucleotides between the linker and the Alu (or truncated Alu) sequences. Since Alu elements preferentially insert into A-T rich regions (Daniels & Deininger, 1985) and *MseI* cuts at the sequence TTAA, then this result is not surprising. The advantage of using *MseI* for the restriction digestion is that most of the Alu-linker products are

small enough to be amplified. Although it would be difficult to perform nested PCR in the opposite direction with those few A-T rich nucleotides, searching GenBank using the BLAST program with the obtained flanking unique DNA sequences as the query may in some cases identify the rest of the genomic sequence for each Alu element. This will provide the Alu location with both its flanking sequences. Flanking unique sequence primers can then be designed and the Alu polymorphism can then be confirmed using other human DNA sources. Once the polymorphism is confirmed subsequent population studies can be performed.

Display and rapid identification of Ya8 associated variants

To alleviate the need for testing every Ya8 element obtained by this assay, we chose to end-label the third round nested PCR primer to enable a display of individual Ya8 repeats following electrophoretic separation and autoradiography. Observed variations may be due to primer mismatch, genomic rearrangements, small insertion/deletions or Alu based insertion/deletions (I/D).

We carried out the procedure with four different individuals to discern which bands represent variants (Figure 5), and to effectively display variants as DNA fingerprints. We obtained about 40 bands per individual from a single reaction. Among the four individuals analyzed, about one half of the bands appeared variant (Figure 5). We have developed a potent method for the generation of Ya8 associated DNA fingerprints that is in reasonable agreement with the database mining approach and seems to display the majority of Alu subfamily members. This necessitated addressing what proportion of the fragments generated were the result of the presence of a Ya8 Alu element and whether the lack of the same band in another individual represented an Alu insertion polymorphism. We chose 12 bands to re-amplify and verify as Ya5/8 elements. Those bands that appeared variant were analyzed for Alu insertion polymorphisms. Other bands were selected for future testing of dimorphisms as these individual Ya8 elements may display variation among other people/populations. Occasionally, upon re-amplification from the isolated band, we obtained background products and therefore, generally more than one clone was sequenced. Of the 12 isolated bands (Figure 5) nine were verified as precisely amplified HS16R-LNP products. Two others each contained

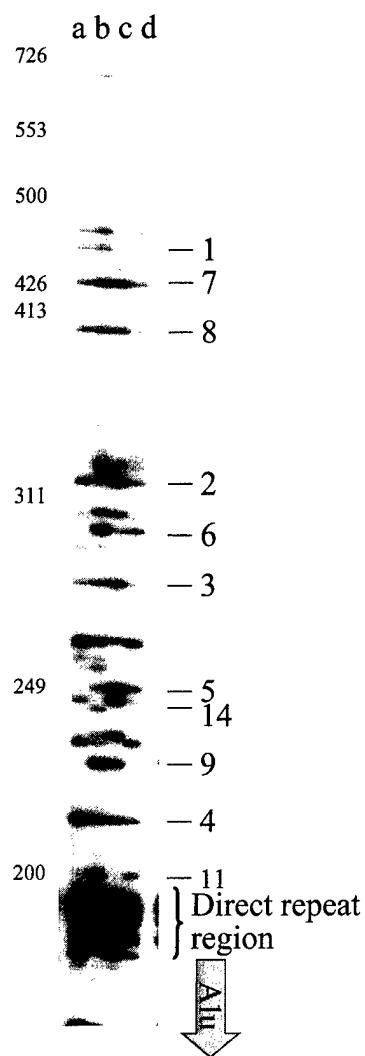


Figure 5. DNA fingerprints of unrelated individuals based on anchored-Alu PCR. Individual bands are numbered for identification purposes. Fragment lengths are shown in nucleotides to the left. DNA samples used are of Caucasian (lane a), Hispanic (lane b), Hindu-Indian (lane c) and Chinese (lane d) descent.

a Ya5/8 Alu, one randomly amplified by HS16R (anc-8) in lieu of the linker primer, while anc-3 contained sequences downstream of HS16R. Anc14 apparently was an amplified J (PS) Alu element (data not shown). Therefore, this demonstrates the majority of the bands visualized on the autoradiograph are AluYa5/8 repeats and most probably Ya8. The numerous bands at about 178 nt coincide with our previous finding that many of the products will have between 12 and 18 unique sequences. Of the nine bands where we attempted to obtain the opposite flank by nested anchored PCR, we reached the opposite (downstream) flank of the Alu for

three of them (anc-5, anc-6, anc-4). In some cases the amount of unique sequence was too small to employ nested primers, and in some cases there was a high level of A-T richness. In one case we merely got a non-specific product. All three sequences obtained were authentic Ya8 Alu elements based on the diagnostic nucleotide positions and the high level of conservation of the sequence in relation to the consensus. This demonstrates the successful nature of our protocol to select for this subfamily of repeats amongst a large background of Alu repeats.

When 'crossing' the anc-5 Alu by nested PCR using four individuals (not all identical to Figure 5), we found a correspondence between the generation of a distinct band among the individuals that also had the anc-5 band on an autoradiograph. However, we obtained a short 3' flank of 12 nucleotides that proved difficult in amplifying DNA from various individuals with unique flanks. It is still possible that this variant represents an I/D event. Besides anc-5, anc-6 also appeared polymorphic on the autoradiograph, although anc-4 did not. However, since we had both flanks, for these Alu elements, we developed primers to rapidly assess various individuals for an insertion variant. For anc-6, one of a few different primer sets worked well, yielding the band of expected size, although also generating a few non-specific bands. However, a band was present for 11 unrelated individuals analyzed (data not shown), including those observed on the autoradiograph, suggesting that the anc6 polymorphism was not the result of an I/D variant. In addition, this band was absent in the chimpanzee, possibly indicating the absence of the Alu or perhaps primer mismatch due to nucleotide divergence. Although anc-4 was not variant on the autoradiograph, we tested 13 individuals of various ethnic backgrounds for an I/D event and observed it to be monomorphic. Although we have not verified any of the displayed variants to be the result of an Alu insertion, this potential remains, as we observed Ya8 elements to be highly polymorphic, and all the bands, but one, analyzed were Ya8 repeats.

Discussion

In this manuscript we present an analysis of the smallest defined subfamily of Alu elements located within the human genome termed Ya8. This subfamily of Alu elements was derived from the Ya5 subfamily of Alu elements. The Ya5 subfamily is composed of approximately 1000 members and has largely integrated into

the human genome sometime after the human-African ape divergence. The main reasons that supported the more recent origin of the Ya8 subfamily are the accumulation of three additional diagnostic mutations as compared to the Ya5 subfamily and the lower copy number for the Ya8 subfamily. It is also important to note that a higher percentage of the Ya8 Alu family members (50%) are polymorphic for insertion presence/absence as compared to only 25% polymorphism in the Yb8 and Ya5 Alu subfamilies. These data also suggest a recent origin for the Alu Ya8 subfamily within the human genome. However, it is still possible that the Ya8 Alu subfamily may have amplified from an allelic variant of the Ya5 subfamily that was not as efficient at mobilization as the Ya5 source gene.

The ability to detect a handful of Alu repeats from the background of several hundred thousand Alu elements in the human genome is impressive. The application of computational biology to the analysis of large multigene families such as Alu repeats offers the potential to address a number of new questions in comparative genomics as an increasing proportion of the human genome is sequenced. Studies of the present, as well as ancient, integration patterns of mobile elements in the human genome may begin to be addressed. In addition, the patterns of diversity generated by the integration of mobile elements into the human genome may be analyzed at a scale that was previously unimaginable. These types of studies will shed new insight into the relationships between different types of mobile elements in the human genome, integration site preferences, impact, and the biological properties of these elements.

The development of the ASAP technique facilitated the display of a subset of Ya8 Alu elements from a large and complex background. The preferential isolation of the young Alu elements, as demonstrated here, enhances the identification of recent Alu insertion events in the genome. We focused our efforts on the smallest known defined subfamily of Alu repeats to best address issues of sensitivity of the display of individual elements. One of the advantages of this technique is its flexibility. Altering the restriction enzyme used for digestion of genomic DNA selects for distinct subsets of Alu elements within a particular subfamily, since this technique preferentially amplifies products that range from 200 and 800 bp in size. In addition, modifications to the ASAP technique, such as the use of a less frequent restriction endonuclease, may allow for a display of subsets of the larger groups of Alu repeats such as Ya5 elements. Alternatively, the

use of primers that select for subfamily 'subgroups' may also be used to reduce the complexity of the resultant display by decreasing the number of PCR products. Although we focused on Ya8 Alu elements due to their low copy number, the young Yb8 Alu subfamily is another alternative for ASAP with an estimated copy number of only 1000 elements (Batzer et al., 1995; Zietkiewicz et al., 1994) and some polymorphic members (Hutchinson et al., 1993; Hammer 1994; Arcot et al., 1998). We have previously demonstrated the isolation of young Alu elements (based on sequence identity to a consensus) using a Yb8 diagnostic primer, and a generic Alu as an anchor in the amplification reaction, that can be profiled with minimal background (Kass, Batzer & Deininger, 1996). It is conceivable that variations on the anchored-Alu PCR technique can be employed to rapidly localize individual elements from all three subfamilies of young Alu elements.

Once the flanking sequences of the young Alu elements are obtained, the PCR strategy can be employed to trace polymorphisms that have resulted from recent Alu insertions and are not yet fixed in human populations. The anchored-Alu PCR approach not only facilitates rapid identification of young elements by displaying the amplification products, but will also increase the potential for selecting only those mobile element fossils that exhibit presence/absence variation. Selection in this manner also shifts the spectrum for new elements toward the elements that are lower frequency and less likely to be held in common between individuals or populations. Therefore, this approach should prove to be quite useful for the ascertainment of mobile element fossils to address questions about more recent human diversifications. In contrast, the identification of mobile element fossils using computational biology affords the opportunity to identify multiple frequency classes of Alu elements that are shared at different geographic levels within the human population.

The ASAP method's strength comes from its ability to isolate a subset of interspersed repeat sequences from different DNA sources and compare them at the same time. In other words, this approach is not limited to Alu elements, but may be used with other SINEs (from other organisms) or even long interspersed elements (LINEs) or for that matter any repeated DNA sequence family that has a defined subfamily structure. A second potential application would be the use of ASAP to monitor genomic instability associated with different forms of cancer by providing a multi-

locus monitoring system. Due to its high flexibility the ASAP technique has an enormous range of potential applications.

Mobile element fossils have proven to be simple powerful tools for tracing the origin of human populations (Perna et al., 1992; Batzer et al., 1994a,b, 1996a; Stoneking et al., 1997). These elements should also prove quite useful to the forensic community as paternity identity testing reagents (Batzer & Deininger, 1991; Novick et al., 1993). Some Alu insertion polymorphisms have been identified by chance (Deininger & Batzer, 1995) while others have been identified by library screening in a directed approach (Batzer & Deininger, 1991; Batzer et al., 1995; Arcot et al., 1995a, b, c; Batzer et al., 1996a; Arcot et al., 1998). Here, we have presented two complementary methods involving computational biology and PCR based displays that will enhance our ability to identify the genomic fossils of recently integrated mobile elements from complex genomes. These approaches will contribute to a new era in biological sciences that will increasingly rely upon informatics/computational biology as well as hard-core bench molecular biology to answer global questions in comparative genomics.

Acknowledgements

AMR, MLC, DHK and SUN contributed equally to this paper. DHK was supported by an Eastern Michigan University Spring/Summer Research Award and a University Research in Excellence Fund. AS was supported by a fellowship from the Academy of Scientific Research, Ministry of Scientific Research and Informatics, Cairo, Egypt. This research was supported by National Institutes of Health RO1 GM45668 to PLD, Department of the Army DAMD17-98-1-8119 to PLD and MAB, and award number 1999-IJ-CX-K009 from the Office of Justice Programs, National Institute of Justice, Department of Justice to MAB. Points of view in this document are those of the authors and do not necessarily represent the official position of the US Department of Justice.

References

- Altschul, S.F., W. Gish, W. Miller, E.W. Myers & D.J. Lipman, 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.
- Arcot, S.S., T.H. Shaikh, J. Kim, L. Bennett, M. Alegria-Hartman, D.O. Nelson, P.L. Deininger & M.A. Batzer, 1995a. Sequence

diversity and chromosomal distribution of 'young' Alu repeats. *Gene* 163: 273-278.

Arcot, S.S., Z. Wang, J.L. Weber, P.L. Deininger & M.A. Batzer, 1995b. *Alu* repeats: A source for the genesis of primate microsatellites. *Genomics* 29: 136-144.

Arcot, S.S., A.W. Adamson, G.W. Risch, J. LaFleur, M.B. Robichaux, J.E. Lamerdin, A.V. Carrano & M.A. Batzer, 1998. High-resolution cartography of recently integrated human chromosome 19-specific *Alu* fossils. *J. Mol. Biol.* 281: 843-856.

Arcot, S.S., J.J. Fontius, P.L. Deininger & M.A. Batzer, 1995c. Identification and analysis of a 'young' polymorphic *Alu* element. *Biochem. Biophys. Acta* 1263: 99-102.

Ausubel, F.M., R. Brent, R.E. Kingston, D.D. Moore, J.G. Seidman, J.A. Smith & K. Struhl, 1996. *Current Protocols in Molecular Biology*, Wiley, Canada.

Batzer, M.A., S.S. Arcot, J.W. Phinney, M. Alegria-Hartman, D.H. Kass, S.M. Milligan, C. Kimpton, P. Gill, M. Hochmeister, P.A. Ioannou, R.J. Herrera, D.A. Boudreau, W.D. Scheer, B.J.B. Keats, P.L. Deininger & M. Stoneking, 1996a. Genetic variation of recent *Alu* insertions in human populations. *J. Mol. Evol.* 42: 22-29.

Batzer, M.A., P.L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, C.M. Rubin, C.W. Schmid, E. Zietkiewicz & E. Zuckerkandl, 1996b. Standardized nomenclature for *Alu* repeats. *J. Mol. Evol.* 42: 3-6.

Batzer, M.A., C.M. Rubin, U. Hellmann-Blumberg, M. Alegria-Hartman, E.P. Leeflang, J.D. Stern, H.A. Bazan, T.H. Shaikh, P.L. Deininger & C.W. Schmid, 1995. Dispersion and insertion polymorphism in two small subfamilies of recently amplified human *Alu* repeats. *J. Mol. Biol.* 247: 418-427.

Batzer, M.A., M. Alegria-Hartman, H. Bazan, D.H. Kass, G. Novick, P.A. Ioannou, D. Boudreau, W.D. Scheer, R.J. Herrera, M. Stoneking & P. Deininger, 1994a. *Alu* repeats as markers for human population genetics. IVth International Symposium on Human Identification, 49-57.

Batzer, M.A., M. Stoneking, M. Alegria-Hartman, H. Bazan, D.H. Kass, T.H. Shaikh, G. Novick & P.A. Ioannou, 1994b. African origin of human-specific polymorphic *Alu* insertions. *Proc. Natl. Acad. Sci., USA* 91: 12288-12292.

Batzer, M.A. & P.L. Deininger, 1991a. A human-specific subfamily of *Alu* sequences. *Genomics* 9: 481-487.

Batzer, M.A., V. Gudi, J.C. Mena, D.W. Foltz, R.J. Herrera & P.L. Deininger, 1991b. Amplification dynamics of human-specific (HS) *Alu* family members. *Nucleic Acids Res.* 19: 3619-3623.

Batzer, M.A., G.E. Kilroy, P.L. Richard, T.H. Shaikh, T.D. Dessel, C.L. Hoppens & P.L. Deininger, 1990. Structure and variability of recently inserted *Alu* family members. *Nucleic Acids Res.* 18: 6793-6798.

Batzer, M.A., C.W. Schmid & P.L. Deininger, 1993. Evolutionary analyses of repetitive DNA sequences. *Methods Enzymol.* 224: 213-232.

Bird, A.P., 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic. Acids. Res.* 8: 1499-1504.

Daniels, G. & P.L. Deininger, 1985. Repeat sequence families derived from mammalian tRNA genes. *Nature* 317: 819-822.

Deininger, P.L. & M.A. Batzer, 1995. SINE master genes and population biology, pp. 43-60 in *The Impact of Short, Interspersed Elements (SINEs) on the Host Genome*, edited by, R. Maraia, R.G. Landes, Georgetown, TX.

Deininger, P.L. & M.A. Batzer, 1993. Evolution of Retroposons, pp. 157-196 in *Evolutionary Biology* edited by M.K. Heckht et al., Plenum Publishing, New York.

Deininger, P.L. & M.A. Batzer, 1999. *Alu* repeats and human disease. *Mol. Genet. Metab.* 67: 183-193.

Hammer, M.F., 1994. A recent insertion of an *Alu* element on the Y chromosome is a useful marker for human population studies. *Mol. Biol. Evol.* 11: 749-761.

Hutchinson, G.B., S.E. Andrew, H. McDonald, Y.P. Goldberg, R. Graham, J.M. Rommens & M.R. Hayden, 1993. An *Alu* element retroposition in two families with Huntington disease defines a new active *Alu* subfamily. *Nucleic. Acids. Res.* 21: 3379-3383.

Jurka, J., P. Klonowski, V. Dagman & P. Pelton, 1996. CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. *Computers and Chemistry* 20(1): 119-122.

Kass, D.H., C. Alemán, M.A. Batzer & P.L. Deininger, 1994. An HS *Alu* insertion caused a factor XIIIIB gene RFLP. *Genetica* 94: 1-8.

Kass, D.H., M.A. Batzer & P.L. Deininger, 1996. Characterization and population diversity of interspersed repeat sequence variants (IRS-morphs). *Genome* 39: 688-696.

Labuda, D. & G. Striker, 1989. Sequence conservation in *Alu* evolution. *Nucleic. Acids. Res.* 17: 2477-2491.

Miyamoto, M.M., J.L. Slightom & M. Goodman, 1987. Phylogenetic relations of human and African apes from DNA sequences in the psi eta-globin region. *Science* 238: 369-373.

Munroe, D.J., M. Haas, E. Bric, T. Whitton, H. Aburatani, K. Hunter, D. Ward & D.E. Housman, 1994. IRE-bubble PCR: a rapid method for efficient and representative amplification of human genomic DNA sequences from complex sources. *Genomics* 19: 506-514.

Novick, G., T. Gonzalez, J. Garrison, C. Novick, M. Batzer, P. Deininger & R. Herrera, 1993. The use of polymorphic *Alu* insertions in human DNA fingerprinting, in pp. 283-291 *DNA Fingerprinting: State of the science*, edited by S.D.J. Pena, R. Chakraborty, J.T. Epplen and A.J. Jeffreys, Birkhauser Verlag, Basel.

Perna, N.T., M.A. Batzer, P.L. Deininger & M. Stoneking, 1992. *Alu* insertion polymorphism: A new type of marker for human population studies. *Human Biology* 64: 641-648.

Shen, M.R., M.A. Batzer & P.L. Deininger, 1991. Evolution of the Master *Alu* Gene(s). *J. Mol. Evol.* 33: 311-320.

Stoneking, M., J.J. Fontius, S.L. Clifford, H. Soodyall, S.S. Arcot, N. Saha, T. Jenkins, M.A. Tahir, P.L. Deininger & M.A. Batzer, 1997. *Alu* insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* 7: 1061-1071.

Zietkiewicz, E., C. Richer, W. Makalowski, J. Jurka & D. Labuda, 1994. A young *Alu* subfamily amplified independently in human and African great apes lineages. *Nucleic. Acids. Res.* 22: 5608-5612.



Large-scale Analysis of the Alu Ya5 and Yb8 Subfamilies and their Contribution to Human Genomic Diversity

Marion L. Carroll¹, Astrid M. Roy-Engel², Son V. Nguyen¹
Abdel-Halim Salem^{1,3}, Erika Vogel², Bethany Vincent^{1,3}, Jeremy Myers^{1,3}
Zahid Ahmed¹, Lan Nguyen¹, Mimi Sammarco¹, W. Scott Watkins⁴
Jurgen Henke⁵, Wojciech Makalowski⁶, Lynn B. Jorde⁴
Prescott L. Deininger¹ and Mark A. Batzer^{1,2*}

¹Departments of Pathology, Genetics
Biochemistry and Molecular Biology
Stanley S. Scott Cancer Center
Neuroscience Center of Excellence,
Louisiana State University Health
Sciences Center, 1901 Perdido Street
New Orleans, Louisiana, 70112, USA

²Tulane Cancer Center, Department of
Environmental Health Sciences
Tulane University Health Sciences
Center, 1430 Tulane Ave., SL-66
New Orleans, Louisiana, 70112, USA

³Department of Biological Sciences
Biological Computation and
Visualization Center, Louisiana State
University, 508 Life Sciences
Building, Baton Rouge, Louisiana
70803, USA

⁴Department of Human Genetics
University of Utah Health Sciences
Center, Salt Lake City, Utah,
84112, USA

⁵Institut fur Blutgruppenforschung
Hohenzollernring 57, D-50501, Köln
Germany

⁶National Center of Biotechnology
Information, National Library of
Medicine, National Institutes of
Health, Bethesda, MD, 20894, USA

⁷Laboratory of Molecular Genetics
Alton Ochsner Medical Foundation
1516 Jefferson Highway, New Orleans
Louisiana, 70121, USA

*Corresponding author

We have utilized computational biology to screen GenBank for the presence of recently integrated Ya5 and Yb8 Alu family members. Our analysis identified 2640 Ya5 Alu family members and 1852 Yb8 Alu family members from the draft sequence of the human genome. We selected a set of 475 of these elements for detailed analyses. Analysis of the DNA sequences from the individual Alu elements revealed a low level of random mutations within both subfamilies consistent with the recent origin of these elements within the human genome. Polymerase chain reaction assays were used to determine the phylogenetic distribution and human genomic variation associated with each Alu repeat. Over 99% of the Ya5 and Yb8 Alu family members were restricted to the human genome and absent from orthologous positions within the genomes of several non-human primates, confirming the recent origin of these Alu subfamilies in the human genome. Approximately 1% of the analyzed Ya5 and Yb8 Alu family members had integrated into previously undefined repeated regions of the human genome. Analysis of mosaic Yb8 elements suggests gene conversion played an important role in generating sequence diversity among these elements. Of the 475 evaluated elements, a total of 106 of the Ya5 and Yb8 Alu family members were polymorphic for insertion presence/absence within the genomes of a diverse array of human populations. The newly identified Alu insertion polymorphisms will be useful tools for the study of human genomic diversity.

© 2001 Academic Press

Keywords: Alu insertion polymorphism; gene conversion;
computational biology

Abbreviations used: myr, million years old.
E-mail address of the corresponding author:
mbatzer@lsu.edu

Introduction

Alu elements are the most abundant Short INterspersed Elements (SINEs), reaching a copy number of over one million in the human genome,¹ making them the mobile element with the highest copy number. Alu repeats compose greater than 10% of the mass of the human genome. Full-length Alu elements are approximately 300 bp in length and commonly found in introns, 3' untranslated regions of genes, and intergenic genomic regions.²⁻⁴ Amplification of Alu elements occurs through the reverse transcription of RNA in a process termed retroposition.⁵ However, Alu elements have no open reading frames, so they are thought to parasitize the required factors for their amplification from Long Interspersed Elements (LINEs).⁶⁻⁸ Although the human genome contains over one million Alu elements, only a few Alu elements, termed "master" or source genes, are retroposition competent.⁹⁻¹³ The crucial factor(s) that determine an Alu as a functional source gene are not fully known. Several factors have been suggested to influence the amplification process, including transcriptional capacity, priming or self-priming for reverse transcription and others.¹⁴

Alu elements first appeared in the primate genomes over 65 million years (myr) ago.¹¹ Since then, the amplification of Alu elements within the human genome has been punctuated, with the current rate being at least 100-fold slower than the initial rate of Alu expansion within primate genomes.¹⁵ Throughout Alu evolution, the source gene(s) accumulated mutations that were incorporated into the new copies made, creating new Alu subfamilies. Therefore, the Alu family is composed of a number of distinct subfamilies characterized by a hierarchical series of mutations that result in a series of subfamilies of different ages.¹⁵⁻²⁰ Of these subfamilies, almost all of the recently integrated Alu elements within the human genome belong to one of several closely related "young" Alu subfamilies: Y, Yc1, Yc2, Ya5, Ya5a2, Ya8, Yb8, and Yb9 with the majority being Ya5 and Yb8 subfamily members.^{9,18,21,22}

The availability of a draft human genomic DNA sequence as a result of the Human Genome Project²³ facilitates the "*in silico*" identification of recently integrated Alu elements from the human genome.^{17,18} This method proves to be less demanding in comparison to older approaches, such as cloning and library screening.^{9,21,24} These recently integrated Alu elements serve as temporal landmarks in the evolution of our genome, and many of them will prove to be useful in the study of human evolution and in the study of the natural history of different regions of the genome. Here, we present an analysis of the human genomic diversity associated with 475 members of the Alu Ya5 and Yb8 subfamilies in the human genome.

Results

Subfamily copy number and sequence diversity

In order to determine the copy number of each subfamily of Alu elements, we searched the draft sequence of the entire human genome for the presence of Alu repeats using oligonucleotide sequences complementary to each of the subfamilies (outlined in the Materials and Methods). Our query of the draft human genome sequence identified 2640 Alu Ya5 subfamily members and 1852 Alu Yb8 subfamily members. Both of these copy numbers are in good agreement with previous estimates of the sizes of these Alu subfamilies based upon high-resolution restriction mapping and computational biology.^{18,21}

A comparison of the nucleotide sequences of all of the Ya5 and Yb8 Alu family members can be found at our website (<http://129.81.225.52>). In order to determine the time of origin for the respective Ya5 and Yb8 subfamilies, we divided the nucleotide substitutions within the elements in each family into those that occurred in CpG dinucleotides and those that occurred in non-CpG positions. The distinction between types of mutations is made because the CpG dinucleotides mutate at a rate that is about ten times faster than non-CpG positions^{9,25} as a result of the deamination of 5-methylcytosine.²⁶ In addition, all insertions, deletions and 5' truncations were excluded from our calculations. A total of 441 non-CpG and 241 CpG mutations occurred within the 231 Alu Ya5 subfamily members used in this analysis. For the 244 Alu Yb8 subfamily members analyzed, a total of 478 non-CpG and 275 CpG mutations were observed. Using a neutral rate of evolution for primate intervening DNA sequences of 0.15% per million years²⁷ and the non-CpG mutation density of 0.799% (441/55,209) within the 231 Ya5 Alu elements yields an estimated age of 5.32 million years for the Ya5 subfamily members. Using only non-CpG mutations in the 244 Yb8 sequences yields an estimate of 5.30 million years old for the Yb8 subfamily (478/60,024). This estimate of age is somewhat higher than the 2.7-4.1 million years previously reported.²¹ However, the previous study of Ya5 and Yb8 Alu family members involved only a small number of elements making the calculated subfamily ages more subject to random statistical fluctuation. Alternatively, the new estimated age based upon non-CpG mutations may be artificially inflated due to sequencing errors in the human draft sequence that may account for an increase in the number of mutations observed.

We can also estimate the ages of each Alu subfamily using CpG-based mutations. The only difference in the estimate is to multiply the CpG mutation density by a mutation rate that is approximately ten times the non-CpG rate as previously described.^{9,25} In this case we calculate an average CpG mutation density for the Ya5 subfamily (241 mutations/11088 CpG bases) or 2.17%,

and (275 mutations/11,224 CpG bases) 2.45% for the Yb8 subfamily. Using a neutral rate of evolution for CpG based sequences of 1.5%/million years yields estimates of 1.44 and 1.63 million years old for the Ya5 and Yb8 Alu subfamilies, respectively. Both estimates are consistent with the initiation of the expansion of the Ya5 and Yb8 Alu subfamilies that is roughly coincident with the divergence of humans and African apes.

Inspection of the nucleotide sequences flanking each Ya5 and Yb8 Alu family member shows that most of the elements are flanked by short perfect direct repeats. The direct repeats range in size from 3-23 nucleotides. The observed direct repeats are fairly typical of recently integrated Alu family members.^{7,9} The appearance of truncations within a number of these elements probably occurred as a result of incomplete reverse transcription or improper integration into the genome rather than by post-integration instability. All of the Ya5 and Yb8 Alu family members analyzed have oligo(dA)-rich tails that range in length from six nucleotides to over 60 nucleotides in length. It is also interesting to note that the 3' oligo(dA)-rich tails of many of the elements have accumulated random mutations beginning the process of the formation of simple sequence repeats of varied sequence complexity. The oligo(dA)-rich tails and middle A-rich regions of Alu elements have previously been shown to serve as nuclei for the genesis of simple sequence repeats.²⁸

Alu Y to Yb8 sequence evolution

In our query of the human genome, we identified 88 Alu elements containing one to seven of the eight Yb8 diagnostic nucleotides. These 88 "mosaic" elements were subdivided into Yb1, Yb2, Yb4, Yb5, Yb6 and Yb7 depending on the number of diagnostic changes present (Figure 1(a)). To facilitate identification of the individual elements with different diagnostic mutation combinations, the mosaic elements were numbered consecutively in order of abundance (Yb1.1, Yb1.2, etc., see Figure 1(a)). No evident sequential order of accumulation of the Yb8 diagnostic mutations can be easily discerned. Interpretation becomes complicated due to the fact that four out the eight diagnostic mutations are CpG changes (positions 1, 2, 4 and 6 Figure 1(a)). The Alu Y has three CpG sites (positions 1, 2 and 6) that become TpG in Yb8, and Alu Yb8 has one (position 4). CpG dinucleotides mutate at a rate that is about 9.2 times faster than non-CpG,^{9,25} as a result of the deamination of 5-methylcytosine.²⁶ Therefore, it is difficult to know if the presence of a TpG diagnostic mutation is due to a change in the Alu source gene or in the particular individual Alu element being evaluated. Because CpG dinucleotides represent hot spots for mutation, a high proportion of CpG positions in the Y subfamily might have mutated to TpG. This makes discrimination between source gene changes and parallel forward mutations occurring in mul-

tiple Y elements at these loci difficult. Therefore, we have eliminated these sites (positions 1, 2 and 6) from our analysis (Figure 1(b)). Position 4 represents a different situation. Because the TpG to CpG mutation occurs at the normal evolutionary rate, it was not eliminated from the analysis. However, some variations may be observed where individual copies might have mutated the position back to a TpG that need to be taken into consideration. Now, a sequential evolution of the appear-

	Y	Diagnostic site								-
		1 CpG	2 CpG	3 C	4 T	5 G	6 d	7 CpG	8 C	
(0)	Yb1.1									d
(1)	Yb1.2							A		
(2)	Yb1.3							C		
(2)	Yb1.4							G		
(3)	Yb1.5							T		
(6)	Yb1.6							T		
(7)	Yb1.7							T		
(12)	Yb1.8							A		
(1)	Yb2.1							T		
(1)	Yb2.2							C		
(1)	Yb2.3							T		
(1)	Yb2.4							C		
(2)	Yb2.5							T		
(3)	Yb2.6							T		
(3)	Yb2.7							A		
(1)	Yb4.1							T		
(1)	Yb4.2							A		
(2)	Yb5.1							T		
(1)	Yb6.1							A		
(4)	Yb6.2							T		
(1)	Yb7.1							A		
(6)	Yb7.2							C		
(12)	Yb7.3							A		
(15)	Yb7.4							T		
	Yb8							G		d
(88)		T	T	A	CpG	A	T	G	d	
		total:	60	36	57	46	22	51	37	35

	Y	Diagnostic site								-
		1 CpG	2 CpG	3 C	4 T	5 G	6 d	7 CpG	8 C	
(2)	Yb1.4									d
(5)	Yb1.2, 7-4.2							A		
(4)	Yb2.1, 3-1.3							C		
(15)	Yb2.6-2.8							A		
(3)	Yb4.1-5.1							C		
(4)	Yb6.2							A		d
(16)	Yb6.1-7.4							C		d
(6)	Yb7.2							A		d
(13)	Yb7.1, 3							A		d
	Yb8	T	T	A	CpG	A	T	G	d	
		#1	#2	#5						#3/4 #3/4

Figure 1. Evolution of the diagnostic nucleotide positions from Y to Yb8 Alu elements. (a) Alignment of the eight Alu Yb8 diagnostic nucleotides and the different Yb1, 2, 3, 4, etc. elements found in the databases. The eight diagnostic nucleotides are indicated in bold at the top for Alu Y, and for Alu Yb8 at the bottom. At position 8, - or d represents the absence or presence of the seven nucleotide duplication, respectively. For easy reference, individual elements containing different combinations of the diagnostic mutations were numbered consecutively in order of abundance (Yb1.1, Yb1.2, etc.). The total number of elements found for each subgroup is indicated on the left in parenthesis. Note that no Yb1.1 was found (0). The total number of the Yb8 individual diagnostic sites found in all the intermediate elements is indicated at the bottom. (b) Alignment of the same elements after eliminating the diagnostic sites in Alu Y elements involving CpG to T changes. Commas separate elements within the same Yb group and dashes between different groups, i.e. Yb1.2, Yb1.7 and Yb4.2 represents Yb1.2, Yb1.7 and Yb4.2. The suggested evolutionary order of the occurrence of the changes at the diagnostic sites are indicated at the bottom (#1, #2...).

ance of the diagnostic sites can be obtained, starting with position 3, then 4, 7 and/or 8, and finally position 5 (Figure 1(b)). The mutation at position 3 appears to have occurred first, being the most common single nucleotide change with 15 Yb8 mosaic elements. The other Alu Yb8 mosaic elements with only one diagnostic nucleotide change occur in lower frequencies and may be explained by parallel mutations, post-transcriptional selection,⁸ or by a forward gene conversion event. The order in which the mutation at positions 7 and 8 (the seven nucleotide duplication) occurred cannot be resolved with these data. Four of the elements (Yb6.2 in Figure 1(b)) do not fit the proposed sequential evolutionary pattern. In this case multiple recombination events would be required to obtain this outcome or some selection occurring at the retroposition process, both highly unlikely. Alternatively, position 5 may be explained by gene conversion events or parallel mutations. The possibility of gene conversion between Alu repeats has been suggested previously.²⁹ In addition, limited amounts of gene conversion between Yb8 Alu elements^{21,30} and extensive levels of short gene conversions in the Ya5 subfamily¹⁸ have been previously reported.

Phylogenetic origin

In order to determine the approximate time of origin of each Alu subfamily member (Ya5 and Yb8) in the primate lineage, we amplified a series of human and non-human primate DNA samples using the polymerase chain reaction (PCR) and the oligonucleotide primers shown in Tables 1 and 2. In this assay, genomes that are homozygous for the presence of an Alu element amplify a PCR product about 400 bases in length. Genomes that do not contain the Alu element at a particular chromosomal location amplify a 100 bp fragment, while heterozygous genomes amplify both fragments. Using this approach we investigated the phylogenetic origin of each Alu element. All 231 Ya5 Alu family members were subjected to this analysis and only one element (Ya5NBC42) was present in the orthologous locus from the common chimpanzee genome. For the Yb8 subfamily, 244 elements were assayed with none being present in the common chimpanzee genome. This suggests that almost all of these Alu elements dispersed within the human genome sometime after the human and African ape divergence and that less than 0.21% (1/475) of the Ya5 and Yb8 Alu subfamily members in the human genome also reside in non-human primate genomes. In fact, this is only the second Ya5 Alu element ever reported that is also found in the genome of a non-human primate.

Human genomic diversity

In order to determine the human genomic variation associated with each of the Ya5 and Yb8 Alu family members, each element was subjected to

PCR amplification (outlined above) on a panel of human DNA samples. The panel was composed of 20 individuals of European origin, 20 African Americans, 20 Greenland Natives or Asians and 20 Egyptians for a total of 80 individuals (160 chromosomes). Using this approach 134 Alu Ya5 (Table 1) and 160 Yb8 (Table 2) subfamily members were monomorphic for the presence of the Alu element, suggesting that these elements integrated in the genome prior to the radiation of extant humans. A total of 28 Ya5 and Yb8 Alu family members appeared heterozygous in all of the individuals that were analyzed, suggesting that they had integrated into previously undefined repeated regions within the human genome as reported previously.³¹ In the PCR-based assay these elements generate a pre-integration site size product from the duplicate copies of the pre-integration site located throughout the genome along with an Alu filled site from the one pre-integration site sequence that contains the new Alu insertion. These elements were not subjected to any further analysis. An additional six elements were located in other repetitive regions of the genome that were identified computationally and discarded from further analysis. The remaining elements were polymorphic for the presence of an Alu repeat within the genomes of the test panel individuals (Tables 3 and 4). Loci that were polymorphic for the presence/absence of individual Alu insertions were subsequently classified as high, low or intermediate frequency insertion polymorphisms (defined in Tables 1 and 2). The unbiased heterozygosity values (corrected for small sample sizes) for these polymorphic Alu insertions were variable, and approached the theoretical maximum of 50% in several cases. This suggests that many of these Alu insertion polymorphisms will make excellent markers for the study of human population genetics. Approximately 25% (58/231) of the randomly identified Ya5 and 20% (48/244) of the Yb8 Alu family members are polymorphic for insertion presence/absence within the human genome. These results are in good agreement with previous estimates of the percentages of insertion polymorphisms within these two Alu subfamilies.²¹

The Alu inserts that have been in the genome longest are more likely to approach fixation. Therefore, we might expect to find different levels of sequence divergence for the Alu elements from each insertion frequency class. Using this approach the average number of non-CpG/CpG-based mutations for the Ya5 Alu family was 1.62/1.06, 2.83/0.67, 2.16/0.66 and 2.53/1.0 for the fixed present, high frequency, intermediate frequency and low frequency Alu insertion polymorphisms, respectively. In the case of the Yb8 subfamily the average number of non-CpG/CpG mutations was 1.86/1.16, 5.0/0.6, 2.2/0.66 and 1.7/1.2 for the fixed present, high frequency, intermediate frequency and low frequency Alu insertion polymorphisms, respectively. In all cases the standard deviations for each average were as large or larger

than the average number of mutations reflecting the heterogeneity in the dataset. No detectable difference in the mutation density within each frequency class of Alu insertions was observed. Therefore, our data suggest that any sequence differences between the polymorphic elements and those with fixed presence may be obscured because of the small number of total mutations and sequencing errors (see Discussion).

Discussion

Alu elements account for more than 10 % of the mass of the human genome. The majority of Alu elements integrated into the genome early in primate evolution. Only a small number of elements (a few thousand) have amplified in the human genome after the divergence of humans and African apes. Here, we report an investigation of the dispersion and insertion polymorphism of the two largest subfamilies of recently integrated Alu repeats within the human genome. Our copy number estimates of 2640 Ya5 and 1852 Yb8 Alu elements within the draft sequence of the human genome are in fairly good agreement with previous estimates of the sizes of these Alu subfamilies although they both exceed the previously published figures.²¹

Using the mutation density and a neutral mutation rate we were able to estimate the ages of each subfamily as 5.32 million years (myr) old for Ya5 and 5.30 myr old for Yb8 using non-CpG-based estimates and 1.44 myr (Ya5) and 1.71 myr (Yb8) using the CpG mutation density. Each of these reported average ages based upon non-CpG mutation density is substantially higher than those reported previously of about 1 myr and 2.7 to 4.1 myr for the Ya5 and Yb8 subfamilies, while the estimates based upon CpG mutation density compare favorably to those previously reported.^{21,32} If we assume a linear amplification of these Alu subfamilies in the human genome, the oldest elements would be no greater than 10.64 myr old for Ya5 and 10.6 myr old for Yb8 using non-CpG mutation density, or 2.88 myr old for Ya5 and 3.42 myr old for Yb8 using the CpG mutation density. The non-CpG based estimates for the oldest subfamily members appears to be somewhat higher than expected for a group of repeated DNA sequences that largely amplified within the human genome after the divergence of humans and African apes which is thought to have occurred within the last 4-6 myr.²⁷ This discrepancy between the two estimates can be explained by considering sequencing errors as a potential factor influencing our current calculations. In the determination of the non-CpG mutations for the estimation of the Alu subfamily age, sequencing errors would be included in the count as mutations, making the estimated age higher than the actual age for the subfamily. If we assume that the sequencing errors are distributed evenly across the entire Alu sequence, then the

number of sequencing errors would be higher in the non-CpG-based estimates than the CpG-based estimates, since there are more non-CpG (242-246) than CpG (only 44-48) nucleotides in the subfamily consensus sequences. Our observation that the levels of sequence divergence from the subfamily consensus sequences do not effectively correlate with polymorphism levels in the human genome also argues that it will not be beneficial to use sequence divergence from the subfamily consensus sequences as a method for the identification of additional polymorphic members of these Alu subfamilies.

We can also compare the calculated ages of each Alu subfamily based upon non-CpG mutation density as a whole to the estimated percentages of Alu insertion polymorphisms and copy number to evaluate the contribution that these elements make to human genomic diversity. Here, we report estimated ages of 1.44 myr for the Ya5 subfamily and 1.71 myr for the Yb8 subfamily. The percentage of Alu insertion polymorphisms in each of the subfamilies was 25 % for the Ya5 subfamily and 20 % for the Yb8 subfamily. The copy numbers of the two subfamilies of Alu elements were also different with 2640 Ya5 Alu elements and 1852 Yb8 elements. When considered together these data indicate that the Ya5 Alu subfamily with both a higher copy number and more insertion polymorphisms has been more successful at amplification within the human genome. In fact, if we assume that the ages of the two subfamilies are about the same the Ya5 subfamily has been about 40 % more efficient at amplification in terms of both copy number and the generation of new Alu insertion polymorphisms within the human genome. Although the sample size is presently small, this is also in good agreement with the number of previously reported Ya5 (six) and Yb8 (three) Alu repeats associated with different human diseases (reviewed in ref. 22). In addition, these data also provide compelling support for the simultaneous expansion of multiple Alu subfamilies within the human genome. The reasons for the differential amplification of the two Alu subfamilies remain unknown. However, they likely reside in the ability of each subfamily to produce RNA for retroposition or at some other point in the process of retroposition itself such as the reverse transcription step. Further experiments will be required to determine the precise molecular mechanism(s) leading to the differential expansion of these two Alu subfamilies within the human genome.

Using the non-CpG-based average ages of the Ya5 and Yb8 Alu subfamilies along with a linear amplification rate we can also estimate the number of members from each Alu subfamily that should be present within the orthologous loci of the non-human primate genomes. Using this approach the oldest Alu repeats from each subfamily would be approximately twice the average age. In other words, the Ya5 subfamily would have begun to expand 10.64 myr ago with the Yb8 subfamily hav-

Table 1. Alu Ya5 accession numbers, locations, human diversity, oligonucleotide primers and PCR parameters

Name	Accession	5' Primer sequence (5'-3')	3' Primer sequence (5'-3')	A.T.*	Human diversity ^a	Chr. ^c Loc.	Product size ^b
Ya5NBC02	M28713	CITTAGACITACAGTTGTTAGCCCTTG	CTGCACTTCCAATTCTTCTACCAC	55	FP	22	710-384
Ya5NBC03	AL023807	ACTGCTTGAAGCTGAACCTAAGAACCC	CTCTGCTCTCTCTAGCTTGTAAATAAC	60	FP	6	564-243
Ya5NBC04	AL008629	CTGATGAAATCTGCTGTATTG	GCAAACTCAAACAGGTTAAACAC	60	R/R	X	483-154
Ya5NBC05	AC007363	TAGGATTAATCTACTAAAGCGTAGATT	GTTTAAGCTTAAGGTTTAAACAAAGAGT	60	IF	2	476-163
Ya5NBC06	AC008344	GATTACATCTGTGATCTGAAACT	GAACATTGTTCTTGTGACTGT	60	FP	6	539-189
Ya5NBC08	AC006478	ATTAAAGCACCAGAAATTGCCATAC	CTAGTAAGGCTAGTGTCCATAATTGAAAGTG	55	FP	7	513-196
Ya5NBC09	AC006582	CITCCCTTAAAGGTTAACCTAACAAAGAC	TITTCACACTTGTACTGAGGACAGGAC	60	FP	Y	415-102
Ya5NBC10	AC0086725	AAAGCTAAAGAAAGTGGCCAC	CAATGAAGATAAGACGCCCTA	60	FP	20	449-141
Ya5NBC12	AC0093076	GIGACTCAAAAGGACTTACACGCTC	CITCCGTCTCCCTAACGTAATT	60	FP	2	470-129
Ya5NBC13	AL031302	CTTCCCTGTGATACCTCTGAC	GTCGTGACCTGCAACAAAG	60	FP	22	604-291
Ya5NBC14	AL050342	Alu flanked by other repeats	-	-	-	1	-
Ya5NBC15	AC008676	CITTCCTCGCTTAACTTCTACTG	GAAGATTAGCTGAGAACAAAGAC	60	-	5	502-192
Ya5NBC16	AC008608	CITTAACCAAATAATGGAGGAGTT	CAGAGATTACTACTGAACTGTGACG	65	HF	5	539-229
Ya5NBC17	AC007076	TATGCTCTGAGGGTTCTAGATCTCTG	GAATAGGAGCATATTCAGTTGAG	65	FP	7	553-232
Ya5NBC18	AC008433	AGACCATCTTAAAGGAAATTACATGT	GAACATGAAAATTGTTAATAAGGTGTC	65	HF	5	495-180
Ya5NBC19	AL109548.1	AATCACTGTACTCATGGGTTCT	AGACCTAGCTGCTCTACTCACTGT	60	FP	1	433-116
Ya5NBC21	AC008482	AGATGTGCTTCAATGAGACTTCT	AAAGCTCACTCATAAAGAACAC	60	FP	5	512-177
Ya5NBC22	AC004519	TCTGTGTTCTGATGTGTTACTCTTA	GAATGTAAGCTGTACTTCTCTTAAT	55	IF	7	471-156
Ya5NBC24	AJ011932	AAATTGAGAGAACGAGGAGGTT	CCTGTAATCAACTGTGTAACGTCAC	60	IF	21	538-286
Ya5NBC25	AC004220	GTGAGGGTAAAGCAAGCTCTCT	CACCAAGCTGTAAGTGTCTCTCA	55	FP	5	538-218
Ya5NBC26	AP003111	GGGCTATTCTGATTCTCTCTCTC	AGAAAGACATACAGACATACTACAGAC	55	FP	21	476-158
Ya5NBC27	AC003691	CTGATAACAGGTTACACTGAAAGAAC	ACAGTGTAAAGCTAACCTACAGGAGAT	55	IF	11	591-265
Ya5NBC28	AC005862	GGGTACATGTGAGGGTTCTTATAC	GCTACTGTGTAACACACAGATACATAG	55	IF	7	474-191
Ya5NBC30	AC007159a	CTGACACATAGTGTGTTCTAAAG	GAGTAGTTGAGCTGTTTGTAGGAG	60	FP	?	502-191
Ya5NBC31	AL033543	GTATCTCTGTTCTTCTCAACAGACTGAG	CTCTATTCTCACTTCTGAGTC	60	FP	22	523-238
Ya5NBC33	AC006288	GCAATTGGTATCTGGGTTCT	CTCTAGCTACAGTTTTCCTCATGTATC	60	FP	9	543-226
Ya5NBC34	AL031575	CAGCTGTACTATCTGTCGCTGTAT	TGAGACATACAAACAGAAATCT	60	FP	X	494-150
Ya5NBC35	AC004534	GAGAGCTTACAGAGGAGCTATT	GTAGCTGATGAGGAAGAAAAGAC	60	IF	7	515-179
Ya5NBC36	AC004066	ATGAAATAACTCTGAGTCGCTC	AGTTCTGTGTTGTTCTTAAATACCT	60	R/R	7	515-200
Ya5NBC37	AC002476	GCTTGGGGTTCTACTCTCTATCTT	ACTGTTATAAGCTTCTCTTCTTCTTC	60	IF	X	497-184
Ya5NBC38	AC0065033	GTACCTCTTAATTACAGCTCTCATCC	GAACCTTCTGGGTGAAAAATCTAG	60	IF	7	487-170
Ya5NBC39	AC0055333	TGGACTCTACGTTTGTGTTCTA	CTAACACAGGTTACGACCTCTT	60	FP	14	469-152
Ya5NBC40	AC008887	ATTGATCTCCACACTGATGCCCTA	GACAAAGACTTACCCCTACTATT	58	-	5	417-105
Ya5NBC41	AC008828	CTCTTATGGACTGACAGCA	GTTCTACATGCTCATATACTGTAGGG	55	FP	5	441-128
Ya5NBC42	AL078621	AGTAAGTCCTCCCTCCATAGCT	GGCTTTCTAACCCAAAGGTC	55	FP	22	485-165
Ya5NBC43	AL096867.7	CCTTCCCTACTAGACAGACACAT	CTTTAAAGCTGCTTGTGTTTGTGTT	55	FP	6	539-218
Ya5NBC44	AL096840	CATTCTCTCTGTTCTCTTCTTAT	GTGAGTTGGGGATATGGTGTGAG	55	FP	1	465-155
Ya5NBC45	AL049868	TAGGGTAAGAATATGTCGCTTGTAG	GTCCTGTAACGACTATGTGAGGAG	60	IF	20	591-255
Ya5NBC46	AC009466	GATGTGAGATACTGTGATGTCAG	GTAAGCTTGTGAGCTGCTGCTCTGT	55	FP	11	405-94
Ya5NBC47	AC007227	CTCAAGATTGGCTATAGTGTATT	AGACACAGGTTACAGTAAAGAGT	55	FP	1	526-200
Ya5NBC48	AC002290	ACTGTTAGATAGTGTATTACTGCTCC	AACCTACAGCTGTTACAAACTACAGA	55	FP	11	481-176
Ya5NBC50	AL096829	AGCCTGGCCCATATAAAA	GTTCATAGATAAGACACTGCTGATTTACT	-	IF	1	465-155
Ya5NBC51	AC008249	ATTTCCGAAAGTTCTTACATCTAGTC	AAAGCTTTAACGCTTACACCTCTCT	60	FP	3	437-140
Ya5NBC52	AC009464	AAGTGAAGTCCAAAGTCATG	CTTCTGCGCAGTAAATGAAAG	53	-	16	688-371
Ya5NBC54	AL024507	GTTTATGTCAGTAGGAGTTCTCGTGTAG	TCATTGTTATCATCTGCTGTACCTGT	60	IF	6	433-130

Y55NBC56	TCATGTTATCAGTCGTCAGCTCTG	AGTCACAAATAGATGTTATTGGAGTCAGG	60
Y55NBC57	GACGTTAAGGAGATGTTGAATGAAAAAT	ACTGTAGGAGGATAATGGAAAGTCACAGA	60
Y55NBC58	TGCCCTTAAACCAATTTCCTTATTTC	TATTGGCTGGATTGTTATCTCTTGG	60
Y55NBC61	TGAAATAATCAGTGGGAAG	GTATATCTTACCGAGACTCAGTTTTCAGC	55
Y55NBC65	ATGGTAAATTCCTGTTGTCAG	GTAATGTCCTCCATTGTCATTG	61
Y55NBC67	AC006210	TGCAATTCTCTGGAGATTGACT	58
Y55NBC68	AC006005	ATGCAAGCAGACATTAGTCAGT	50
Y55NBC69	AC004053	TTCAAGAGGACACATTAGTCGCT	60
Y55NBC70	AC004454b	TATGACTAATGTTGGGGCTT	60
Y55NBC72	H5234H5	TTTACTCTGTTCTCTGTTGTC	51
Y55NBC73	AC004454a	GGTGGAGGAAAGTGGAGAGAA	60
Y55NBC73	M95868	GGGCAACCTCAACAGGAT	60
Y55NBC76	AL008629	TGGTGGGCTAGTATTGAGTGA	60
Y55NBC77	AC006155	CTGGCATGGGATTCTTCAC	60
Y55NBC78	AC004454b	ACACTAACATCCATCACCTT	60
Y55NBC79	H596017	CGTGGCATGGAGATGAGATGAAA	60
Y55NBC81	H595742	AGTGTGGAGGAGGAGT	60
Y55NBC82	AC005217	AGGAAAGTCGCTGGACAC	35
Y55NBC83	AL022101	CCTCCCTAGGAGAGACAT	20
Y55NBC87	AL009830	CACAGGATTAGAGAGATGATGATGATAG	20
Y55NBC89	AC009807	GAGGCCAAAATTCCTTACA	11
Y55NBC91	AL034378.2	GTATTTAAACCCATAGGCAACCAT	11
Y55NBC93	AC010068.1	TAGGAGGAAAGTGGAGGAAAC	55
Y55NBC94	AC008788.1	CACTTAGGATGCTTATCTTAC	55
Y55NBC95	AC009862	GAATGAACTCATGGGATATGTTCT	57
Y55NBC96	AC004547	GTATTTGAAACCCATAGGCAACCAT	60
Y55NBC97	AC004453	TGTGAGTTAGAAGCAGTAAAGAG	60
Y55NBC98	AL049591	CTGTGAGTTAGTACTGTTGGACCTATG	35
Y55NBC98	AL031312	CCIGACTCGAAAGTCTGTTTAAAG	62
Y55NBC100	AL005683	TAATATACAGTTGTCCTAGCATACC	55
Y55NBC101	AC006030	ATTACATGAGGAAAGGGAACTGT	50
Y55NBC102	AF118569	CCCATAGAGCTCTTATTTTC	35
Y55NBC103	AL034408	GTAAGCTTGGAGTCAAGGAGAGA	62
Y55NBC104	AC007065	AGAAGAATAGAGGAGACTGTTGTC	55
Y55NBC105	AC008040	TAATACATGAGGAAAGGGAACTGT	20
Y55NBC106	AC005532	CACCCAGCCATTTCAGTAT	20
Y55NBC107	AC004884	CTAAAATACAGGATGAAAGACTGGTAG	17
Y55NBC108	AC007092	GTAAGCTTGGAGTCAAGGAGAGA	55
Y55NBC109	AC005745	AGAAGAATAGAGGAGACTGTTGTC	12
Y55NBC110	AC004761	CTAGAAGGATGAGGAACTCTCT	58
Y55NBC112	AC008032	GGAGGATGAGGAAAGGAACTGT	62
Y55NBC114	AC007782	GTTTATGTTAGCAGATAGTATGTTACAG	60
Y55NBC115	AC009316	GGAAACAAACTGCACTGAGTAAGT	60
Y55NBC116	AC006344.2	GATCCTGAAACTATTTAAAGAAGAC	60
Y55NBC117	AJ010770	GGGGAGGAAAGGAAACACTAGT	60
Y55NBC118	AC005913	AATACGTTGTCCTGTTATGTT	60
Y55NBC119	AC006002	TGTTAAATAACAGAAACACTTCAG	60
Y55NBC120	AC005863	GGACCCACATGACTGAGTAAGT	55
Y55NBC121	AJ011932	AGGGGAAACATCAAAACTC	60
Y55NBC122	AC005747	CCATTICATTCTATTGGGAATTTAG	21
Y55NBC123	AC005739	ATCAAGTGGCAACTGATTACACC	17
Y55NBC124	AL022310	CTAGACAGTGGCAACAGTCTTAATACAG	5
		CATAATGGAAAAACTCATGTCGAC	1
		60	457-131

Table 1. (continued)

Name	Accession	5' Primer sequence (5'-3')	3' Primer sequence (5'-3')	A.T.*	Human diversity ^b	Chr. ^c	Product size ^b
						Loc.	Filled-empty
Y55NBC125	AC004206	AGTATTGTCACCTCTCTAAAGGTGTC	CTGGCTTATGTTTCACTGGATTG	60	FP	6	507-223
Y55NBC126	AC005144	GTCGCTTAACTGATTAACAAACAC	GTGGCATTCTACTACTGAAACCTAAC	60	FP	17	480-171
Y55NBC128	AC004808	GGTGGGACAAAGAAATCAAC	GCTTATGGCTGTCACCYCAAC	55	FP	7	648-293
Y55NBC129	AL008635	TACATGGATTAGGCCGCTCA	ACAGTTTACTGACTACACAGGATTTCAG	60	FP	22	486-180
Y55NBC130	AC004629	GTTGTCACCTTTGACTAGTATGA	GACAGTTTACTGACTACACAGGATTTCAG	60	FP	5	602-287
Y55NBC131	AF02896	CCCAAGATCTAGGTGATGACAC	GCACCTGAGATAACCTAGTTGATGTC	60	IF	X	495-174
Y55NBC132	U91328	CTCGTGAATCTAGGTGATGACAC	CGGGTTGATCCTTAAATACATCAT	60	IF	6	458-228
Y55NBC133	AC000355	TGTTATCATACACAAATACGACCTTAA	TCTTGGCTATAGGATATGAAACTTAAC	60	FP	7	692-274
Y55NBC135	U01102	ATTAAGCTATGGTAAACGGCAC	GACTCTCTCTGTTGATTAGAAACAG	60	LF	11	436-117
Y55NBC136	AC008124	CAGAACATCATCAAGTTATAATGC	GGAAATTGAAATGATGGCAA	60	FP	12	749-439
Y55NBC137	AC005002	GTGCTGTTCTGTCGTCAC	GCATAGAGACCAATCTGGAG	55	RR	?	521-197
Y55NBC139	AL031650	TGAAAGCTTAAGGTCCTCTCTCT	TATGAGCAGAACAGGGAGACAG	60	FP	20	851-534
Y55NBC140	AC007877	GCAGGCCAAGTGTAAATCTAT	GGTGTGTTAATGTCATCATAAACG	60	FP	2	471-135
Y55NBC141	AL096769	CTGAGAACCCACAAAGTACTGAC	CATGGACCCATATACAGTACACAA	60	RR	20	480-139
Y55NBC142	AC007392	ACATTCTGACACCTGTCAGTCAT	GGTGTGTTAATGAAAGAAATATC	60	FP	2	663-321
Y55NBC143	AC006374	GCAATGACATATAGATGTCCTC	CTTTCCTACCATGTTGTCCTT	60	FP	7	572-251
Y55NBC145	AL035657	TGCACTCCCTCTGTCGTC	ATTGGGTTTACTAGACAAAGG	60	FP	20	500-276
Y55NBC146	AL022329	CTGTCCTCTCAGGTCACTT	CTAGCATGTTGTCACCTCTAAC	53	RR	22	604-131
Y55NBC147	AC007856	TAAGGGTGGACAGATGAAATAAAC	AAATATCACCTTATCAGTGGGACCT	60	LF	12	493-155
Y55NBC148	AL031659	ACAAAGTACACATGTAACCCAC	AAGGTGTTGTCAGACTATCTCT	60	IF	20	505-193
Y55NBC149	AL033525	GTTGTTACTGTGCCAACTATCTCAT	ACTTATATGACGGGGTACAGTTCT	60	FP	1	466-156
Y55NBC150	AF135628	AAATGGGACACAGGGTGTAAAGA	CCAAACTGCTATTTAAAGGGTAG	60	IF	19	491-169
Y55NBC152	AC004953	Alu flanked by other repeats	Alu flanked by other repeats	-	-	7	-
Y55NBC153	AC005820	CCAATCTGGAAATTATGACAGTAG	CTTGACTCTCTGCTGTTCTCT	60	FP	7	496-186
Y55NBC154	AC006371(B)	AAACACCCCTAGATGGTAA	AGATGAGTGGCTCAGAACAAAG	60	IF	197	501-197
Y55NBC155	AC006365	TGTCATATATCAGACAGATCATGAG	ACTTCAACTATGTTGTCAGTTTG	60	IF	182	503-182
Y55NBC156	AC002531	TGTGTTAAGTGTAGTTCAAAAGGTT	GGTGTGACTGAGTAAACATAAA	55	IF	17	551-234
Y55NBC157	AC005281	CATACTGTAATACACTGTTACTCA	TCAGAAAGTATACAGTGTGATGCC	60	IF	17	480-148
Y55NBC158	AC005019	TATCCCTCCATACCAAATCTTC	GGATGGATTAGAAAGGGTAGATTAG	60	FP	7	516-207
Y55NBC160	AC005245	CTCACTGCTGCTGATACTCTATA	GCCATCTGGCTAAATCACACATT	55	IF	17	500-172
Y55NBC161	AL031978	CCTGTTAAGTCAACTGAGAAGAAGA	GCAATGAGAAAGTACACGGCTCTAA	60	FP	6	491-199
Y55NBC162	AC003957	ATGAGCAAGTCTCATATCTCTCA	CTTGTGTCGTTCAAGGGCTATA	60	FP	17	481-167
Y55NBC163	AC004057	CAAACAGAAGTCTTACCACTAG	TAGTAAAGTACACGGCAAGA	60	FP	4	624-316
Y55NBC164	AF042090	CTGCTGACTCTGAACTTAACCTC	TAGGGAAAGTGTCTTAGGGTCTCT	60	FP	21	503-190
Y55NBC165	AC004040	CCCTTGCTCTATAGATAAGTTGG	ACTGGACAAAACTAGAGGAGAA	60	FP	1	451-130
Y55NBC167	AC003980	AGGCCACGCTTAACGTTACTA	GTGGGTCTTAAAGTTTCAAT	60	FP	7	473-155
Y55NBC168	Z97876	AGTGTCAACCGAGATGTGTGAC	TGCTACAGTCCTGACTTCA	53	-	16	515-239
Y55NBC169	AC002456	TATAATCCACAGTAAACCCCTCA	GCTATAGCTCTTATTTACGGGAAT	45	FP	1	492-164
Y55NBC170	Z94722	GCAAGACCTGTTGACGTTAAAT	ATAGTTGATACAGCCAAAGACA	60	FP	7	493-184
Y55NBC171	AL035688	TCTGAATTAACAGTCAGCCATC	GAGAGTACAGCAAATACAGGCC	60	FP	6	521-195
Y55NBC172	AC006371	CCAAACGTAAGATTGAGTGG	CTTCCTATCCCTGCTACATAACAT	55	IF	130	451-130
Y55NBC173	AC003977	AGCACACGAACTGGATTAAT	AGTGTGTTCTGCTGTTATTTC	55	IF	13	666-298
Y55NBC174	AC006462	TCACTCTTGTCTGCTGACTAG	TGCTACAGTCCTGACTTCA	53	-	16	473-155
Y55NBC175	AC008396	CCAGTGTACAGTCAGTAAAT	GCTATAGCTCTTATTTACGGGAAT	55	FP	107	508-107
Y55NBC176	Z74739	GGGGGACTATGGTTGATATAC	GGACATGGCTCAGGTTATTTC	55	FP	14	526-206
Y55NBC177	AC000111	TTTCCAAAGGGTCCCTAT	TCATGGAAAAGCTGTTGGT	53	R/R	7	617-300
Y55NBC178	AC004900	AGAGCTGGTGAACCTGTTGAGAC	GAGCCATGATAGGGAAATACTAGAGC	60	FP	7	583-250
Y55NBC179	AC006373	GCAGAAGTCAACTGCAATAGTAGA	GCTGAAACCTTAAACACTGCTGAC	60	FP	20	797-490
Y55NBC180	AL096118	CTTGAAGATGCCATAGTAGA	GGCAATTCTTCTGCTGTC	55	FP	3	525-211
Y55NBC181	AC008041	GTTACAGGAGCTACTCTGGTCTC	AGCCTTCCATCTCATAGACC	58	FP	450-205	

Y45NBC182	AC006365	GAAGGACTATGAGGATGCAAGAAGTC	7	563-287
Y45NBC183	AC005365b	GGACAGCTTAAAGAGGATTTCCTAGA	7	722-410
Y45NBC184	AC000047	CTTGGATAGGCTGGGGCTTAA	9	522-205
Y45NBC185	AC006552	GAGTTTATTGGCTAGGTAGTC	4	513-202
Y45NBC186	AL03545	CACCTCTGAACTAGGGATGCTTAA	6	649-381
Y45NBC188	AC004970	GACAGGACAGATGTTAGGAAAT	7	476-156
Y45NBC191	AC007191	TGACGGGTGAGATGTTATAGAAGC	19	645-330
Y45NBC192	AC005678	CACTCAGATAAAAGATGGGACTTCA	6	536-238
Y45NBC193	AC005065	GTTCATTTCTCTAAATGCCCTCTC	7	525-206
Y45NBC194	AC004966	TATTCCTATGGCGTTATGGCTCTAG	5	486-169
Y45NBC196	AL031785	Alu flanked by other repeats	6	461-204
Y45NBC197	AL031785b	CAGAACTGAGATTGCTGGATGTTGAT	6	450-230
Y45NBC198	AC004603	GCATACTCTTAACTCCATATTCC	4	445-128
Y45NBC199	AC005293	CTTACCATATAACTGGAACTAGA	12	500-200
Y45NBC200	AC005161	GTTAAATGGAAACCAACAGTAG	7	500-174
Y45NBC201	AC004745	CGCCACTTCCCAGTTA	7	499-188
Y45NBC202	AC004603	ACGCTCAAGTCTCACCTCACAT	19	487-154
Y45NBC203	AC004593	GATCAATGACACCAACTCCATCTT	7	445-128
Y45NBC204	AC002385	ATTACAGAGGCTGCCATGAT	7	5199-196
Y45NBC210	AC004848	AGGCCTGCTAGTTGAGGATAATA	7	750-424
Y45NBC212	AC002074	ATCCCTCCAAAGGAGGAA	7	502-190
Y45NBC213	AL078463	TGGAGATGAGCTTCACTTACCT	1	397-91
Y45NBC214	AC004948	CAGCCTGTAAAGCTGGAAAG	7	500-170
Y45NBC215	AL096710	AAGCAAAATCAGCTTACCA	6	780-467
Y45NBC216	AC007245	GAGGGGCTAAGGATAGCATT	7	456-141
Y45NBC217	AC007298	CATTGCGCAAGTGGTATT	12	623-308
Y45NBC218	AC006969	AGGAGAGTGGAAATGGGTTA	Y	445-109
Y45NBC219	AC005989b	TGTTGTTGAAAGGAGGAA	Y	445-129
Y45NBC221	AC004019	GCCAACTAAAGGAATAATCA	22	640-313
Y45NBC223	AC005006	GATGTGACCCCTGGGTTAA	22	455-214
Y45NBC224	BK407F11	TCCAAACCTTTTGTCTG	22	502-195
Y45NBC225	AC002470	AGCCAACTATCTGGTTGTTG	22	421-130
Y45NBC226	DJ323M22	CCTGCAACACCACTTAC	22	842-529
Y45NBC228	AC004632	ACTGATGCGCAGGCCTCA	-	-
Y45NBC229	AL096873	Alu flanked by other repeats	22	-
Y45NBC230	AC000100	GACAAAGAAATGTCAACAGGTAA	19	418-99
Y45NBC231	BA422A16	Alu flanked by other repeats	22	-
Y45NBC231	AC007682	TCACTGCTGAACTCTGAT	55	795-677
Y45NBC232	AL036565	CCTGATACCCACACATACC	55	395-72
Y45NBC233	AL0136295	CTCCCAAGGTCCTGATGTC	14	426-111
Y45NBC234	AL132642	GAGCTACTGGCACCTCCAC	14	368-60
Y45NBC235	AP000866	CACATGGACCTGTTGCTGT	21	493-190
Y45NBC237	AL133289	TCCCTGAAACAAACCCATT	65	450-144
Y45NBC238	AL133404	CAACAGAGGAGAAATGATCAGTGG	6	429-147
Y45NBC239	AC020663	CCTCTACCTGGCTGGTTCAA	16	425-114
Y45NBC230	AC008372	ATGCAAATGCGCGATGTC	14	535-207
Y45NBC231	AC008943	TCTTGGCAAGGAGATGTCAA	5	564-279
Y45NBC232	AC010659	CACTCAGCATCCAGTTCACG	3	385-54
Y45NBC233	AL121823	CACTTGGCAATTCAGTCCAAA	X	481-174
Y45NBC234	AC016025	GTGGGGCACTGTGATTC	22	392-70

Table 1. (continued)

Name	Accession	5' Primer sequence (5'-3')	3' Primer sequence (5'-3')	A.T.*	Human diversity ^b	Chr. ^c Loc.	Product size ^b
Y56NBC315	AP000474	GTAGACACGGCAGGCAACTC	AAAGGATCGTAAGAGGAGA	62	FP	21	444-134
Y56NBC317	AL132865	CCAACTGCGCCACAAATAG	GATGGTAACTTCTTCTGGT	60	FP	14	384-64
Y56NBC319	AC007395	TTGCTGGTCACAAACCAATA	CCTTGTGATCATGGTGCTG	60	FP	2	368-77
Y56NBC320	AC009498	CCATTCCTCAATTGTCATA	CCATGGGAGAGGTTCAAA	60	FP	2	478-161
Y56NBC321	AL121748	GGGATCTCTTTCACCAA	GGAGGTGTCATCGGTACA	60	FP	10	455-145
Y56NBC322	AL132860	AGTGCCTAGATCTGTICA	GGGTCTTGAAGAGTCATGG	60	FP	14	451-129
Y56NBC323	AC007076	TTGAAAGGAAGGCCAAAGA	TCTCTGTCCTCCAAACTCTTC	60	FP	7	558-288
Y56NBC324	AC008268	TGTCIAAGGAAATGCAAA	TCCCATGCTCAACTCTTCTT	60	IF	2	486-164
Y56NBC325	AC009479	CTTCTCTCTGAAATGCAAT	CAGTGTAAAGTTGACAATACCC	60	IF	Y	501-184
Y56NBC326	AL133560	CCAAGAGACCACTCCATTTCA	AATGGGGAGAGGAGAGCT	60	FP	X	539-216
Y56NBC327	AL132799	AGGCAGGTTCAATGTTCAA	TTGCTTATGTCGCTGGTAA	60	IF	6	668-339
Y56NBC329	AL121892	TTTTCCCTGTAGTGGAGCA	TGTTCAAGGAGGGAGAGGA	60	FP	20	465-154
Y56NBC330	AL133599	ATGCTGTGGATTCTAAGGA	CTGTCCTGTTGGCTGTG	60	FP	11	402-88
Y56NBC331	AL121593	TTCATGGCAAGGCTGTATA	AGCTCTGGCAGATTAAACA	62	FP	20	414-92
Y56NBC332	AL050342	TGGAAACAGAACATGGACA	ACACAGGTCTGAAATGAGCC	65	FP	1	631-146
Y56NBC333	AL117356	GGCATGTTCAATTCCCAA	CCAACTCTGTTGAGAGTACG	60	IF	14	588-281
Y56NBC334	AL132708	ACATCTGTTGGGACATTC	CCTCGATCCCACTAACATGA	60	FP	14	435-117
Y56NBC336	AC007151	AGGCCACATCACTGTAGG	TGATCATAGCTCTTGTG	60	FP	16	486-172
Y56NBC338	AC009510	TCAAGAAGTTAAAGGACACAA	AGGGAGAGGAAAAGATGTC	60	FP	12	564-271
Y56NBC340	AL109895	TCCATATCCTCTGTGGT	CCTGACCAAGTCCAAATGAC	60	FP	14	468-145
Y56NBC341	AC007899	ATGCAATTGCTGACACAG	GTTGGACCGAGATTCTTCTC	60	FP	2	434-174
Y56NBC342	AL049823	TTTCACAAATGGCACTGA	TGTCCTGTCGCTGTCATTTC	60	R/R	6	604-285
Y56NBC343	AC005660	GACCACATGGTGGGACAT	CCCTCTTGGCTGGTGTG	60	FP	10	457-154
Y56NBC344	AL109853	CGTGAAGAACATAGGCAAC	TCCCTTCCTTATGCTGCAAA	60	FP	Xq	472-158
Y56NBC346	AL096776	GGAGAACATAGTGGGAGAG	ACATCCCCGTGTCATTCCT	60	-	1	396-60
Y56NBC347	AL035411	CATGCCATATGCTTACGTT	TGGGGTAGATTGGACTCATCC	60	IF	1	465-140
Y56NBC349	AC011504	TCAGAACCTGGCAAAAT	GTTGGACCGAGCAGCAGCT	60	HF	19	469-53
Y56NBC351	AP000459	TTCCCTCCGCTTCTCTGT	TGTCACTGTAACCAACCATGCT	55	IF	21	437-123
Y56NBC353	AL034549	CCATGTAACCTGGTAGACGCTT	GTTCAAGGGAAACAGTGTG	60	FP	20	432-119
Y56NBC354	AC008039	GTAGCTGGCCTGTSCTTT	CCTCTGGGCTGAGAAACTCT	65	IF	7	486-148
Y56NBC355	AL078477	CATCTCACATGAAAGCCATT	TGTGCTCTTAATGACCCGGAA	60	FP	11	802-481
Y56NBC356	AF130343	CAGGGCTCTGTAATCAAT	GGAGACAGAGAAAAGGGAGA	62	FP	8	389-84
Y56NBC359	AC007564	GCAAGCTCTATGCAAGGCTAA	AGGCTTTCAAGGCCAGTGT	60	FP	12	775-457
Y56NBC360	AL031121	GAACAAACATTGGTAATGATGC	GACCAATGTCACTTATGAAATCCTT	60	FP	6	407-61
Y56NBC361	AC007270	ATAATTCTCCATTCTTCTG	TGTAAAAGGCAAGGTACAA	60	IF	7	423-131
Y56NBC362	AL050308	CAAGTTGTTGGCATAGGGT	ATCAATCCAGGAGCGTITTT	60	R/R	X	506-187

* Amplification of each locus required 2:30 minutes at 94 °C initial denaturing, and 32 cycles for one minute at 94 °C, one minute at annealing temperature (A.T.), and one minute elongation at 72 °C. A final extension time of ten minutes at 72 °C was also used.

^b Allele frequency was classified as: fixed present (FP), low (LF), intermediate (IF), or high frequency (HF) insertion polymorphism. Fixed present: every individual tested had the Alu element in both chromosomes. Low frequency insertion polymorphism: the absence of the element from all individuals tested, except for one or two homozygous or heterozygous individuals. Intermediate frequency insertion polymorphism: the Alu element is present in all individuals in the populations tested, except for one or two heterozygous or absent individuals. (-) Intermediate frequency insertion polymorphism: the element is present in all individuals in the populations tested, except for one or two heterozygous or absent individuals. (R/R) Repeat in repeat.

^c Chromosomal location determined from Accession information or by PCR analysis of NIGMS monochromosomal hybrid cell line DNA samples.

^a Empty product sizes calculated by removing the Alu element and one direct repeat from the filled sites that were identified.

ing expanded about 10.6 myr ago. If we assume that humans and African apes diverged from each other only 4 myr ago, then we can calculate that 6.64/10.64 (62%) and 6.6/10.6 (62%) of the Ya5 and Yb8 Alu elements should also be found at orthologous positions within the genomes of non-human primates. If we shift the divergence of humans and African apes to 6 million years ago then the estimates change to 4.64/10.64 (44%) and 4.6/10.6 (43%). However, less than 0.21% of the elements were also located in orthologous positions in the genome of the common chimpanzee. The observed distribution of Ya5 and Yb8 Alu repeats located within the common chimpanzee genome would require a human and non-human primate divergence of greater than 10 myr ago. This is clearly a much older divergence time than is commonly accepted.

Three potential explanations may account for this. One is the selective removal of Alu elements from orthologous positions in non-human primate genomes effectively resulting in an ascertainment bias against elements in the non-human primate genomes because our elements were obtained by scanning a database of human genomic sequences. However, we consider this to be highly unlikely, because there are no known mechanisms to specifically remove Alu elements from primate genomes and even when an element is partially deleted from the genome it leaves behind a signature of itself.³³ A second and more likely explanation is that the amplification rate for these subfamilies has increased recently in the human lineage. Alternatively, the higher average ages for each of the Alu subfamilies than those previously reported may reflect a higher sequencing error rate in the genome database, resulting in an inflated age estimate for the Alu subfamilies. The estimated ages of the subfamilies are also inflated by the faster accumulation of non-CpG based mutations (as a result of the larger number of potential target sites) as compared to CpG nucleotides. Therefore, the use of the CpG-based mutation density for Alu subfamily age estimates will be much more accurate than the use of non-CpG mutation density-based estimates using the current draft sequence of the human genome. The magnitude of the putative sequencing errors can be estimated by comparing the previously reported non-CpG mutation density for these Alu subfamilies of approximately 0.4% for the Ya5 and Yb8 Alu elements to the levels reported here of approximately 0.8% for the same subfamilies. Therefore, the maximum possible error rate would be estimated as $0.8\% - 0.4\% = 0.4\%$. In our data analysis, there are a few Alu elements with much higher mutation densities than previously seen. We are not sure whether these represent a small number of authentic, highly divergent subfamily members (approximately 10% divergence), or the concentration of sequence errors in a few elements. Thus, other than the possibility of a few areas where errors may be concentrated, there is a relatively

low sequencing error rate across the entire database, demonstrating the reliability of the draft human genomic sequence. Large scale re-sequencing of the Alu elements characterized in this paper would resolve this issue and allow for an accurate estimate of sequencing error rates within the draft human genomic sequence; it would also provide a refined estimation of the average age of the Alu Ya5 and Yb8 subfamilies as well.

SINE retroposition is the primary mode of mobilization of Alu elements, where mutations in the source gene(s) create their sequence evolution. However, previously we reported that gene conversion and genetic instability might have also significantly impacted the Alu sequence architecture.¹⁸ Our analysis of the Yb8 mosaic elements also suggests that gene conversion may have influenced the evolution of the Yb8 Alu subfamily. Among the alternative explanations for the occurrence of mosaic elements, multiple parallel mutations seems unlikely; unless there was selection for these specific mutations, such as the post-transcriptional selection previously proposed.⁸ However, a selection process that would only select for these specific mutations would be improbable. Recombination may have generated some of these mosaic elements, but multiple recombination events would be required, making it unlikely. Therefore, we believe gene conversion to be the most likely explanation for the existence of the mosaic Alu elements.

Our analysis of the human genomic diversity associated with the Ya5 and Yb8 Alu elements reported here resulted in the recovery of 106 new Alu insertion polymorphisms. The percentages of Alu insertion polymorphisms recovered from each subfamily were 25% and 20% for the Ya5 and Yb8 subfamilies, respectively. The percentages of Alu insertion polymorphisms in these two subfamilies are in good agreement with previously published insertion polymorphism estimates for these Alu subfamilies.²¹ We can also estimate the total number of Alu insertion polymorphisms within the draft sequence of the human genome using our copy number estimates and the percentage of Alu insertion polymorphisms associated with each family. Using this approach we should recover 2640×0.25 or about 660 Ya5 Alu insertion polymorphisms and 1852×0.20 or about 370 Yb8 Alu insertion polymorphisms through the exhaustive analysis of the draft sequence of the human genome. Therefore, the exhaustive analysis of the entire Ya5 and Yb8 Alu subfamilies from the draft sequence of the human genome should generate a little more than 1000 Alu insertion polymorphisms from these subfamilies.

Additional Alu insertion polymorphisms that are present in diverse human genomes may also be recovered using PCR based display approaches such as those previously reported for Alu and LINE elements.^{17,34} Each of the Alu insertion polymorphisms in the genome is a temporal genomic fossil that is identical by descent with a known

Table 2. Alu Yb8 accession numbers, locations, human diversity, oligonucleotide primers and PCR parameters

Name	Accession	5' Primer sequence (5'-3')	3' Primer sequence (5'-3')	A.T.*	Human diversity ^b	Chr. ^c	Loc.	Product size ^b	Filled-empty
Yb8NBC1	AL049798	TACCAAGAGGATGAAAAACAAAG	GGAAACCCAAAGGCTTATAATTAGC	60	FP	1	495-174		
Yb8NBC2	U91327	GTAACTCTCTGGCTCTATATTACT	ATTTTGGATTCAGCATGAACTAACATG	60	FP	12	461-130		
Yb8NBC3	AC004804	AGATGGCGCAAGTCCGATATA	AAAACCTAAAGGGGCGAT	61	FP	12	558-236		
Yb8NBC4	AC005156	AGTGAATGGTGGTGCACA	TGATGCAAGGTTCTGGTC	60	FP	7	461-133		
Yb8NBC5	AC004027	AAGGCTCTAAGCGCACTGGAA	TAAGCTGACAGATGGTGG	60	FP	7	503-167		
Yb8NBC6	AC006150	AT88AAAGACATACACTGG	CITAGAAGGAAATCCAGGAG	60	FP	7	606-203		
Yb8NBC7	AC005148	AGTTTCACTCACTGGTGGAG	CCAACTAGAGAAACGGAGAA	60	FP	7	605-285		
Yb8NBC8	Z98950	AGAAAACTGATGGGGAAAG	TGCTCAAGTCACAGCTA	55	IF	14	599-198		
Yb8NBC9	AC004825	GTCCGCCACCAATCCTATCT	TTCCCTTCAAGGACGTGGA	60	IF	7	685-322		
Yb8NBC10	AC006352	CAGAACAAACGCTTCA	ACCTGAGGAGGACATTC	60	FP	6	505-165		
Yb8NBC11	AL02477	TGAGAAAGTGGTGCCTTAAT	ACCTGAGGAGGACATTC	60	FP	6	510-188		
Yb8NBC12	AC006241	CCAGCAGGGTTTATCTT	ACCTGAGGAAATGTCCTAGTGC	60	FP	9	467-160		
Yb8NBC13	AC002331	TCTGGGTTCTCTGGTGAC	CTGGCAAATGTCACCAAGT	60	IF	16	510-168		
Yb8NBC14	AL08633	AGAGAGAAATTACAACTGACTGC	TGGCTTACATTGGCTATGG	58	FP	X	499-167		
Yb8NBC15	Z82211	CTAACTCCCTCTGGCTGATA	CTCTGGTACGGCATAAAGC	60	FP	X	481-165		
Yb8NBC16	U51244	TACCGAGGGCCACTTGG	CGGACCAAGCAGCAGTAT	60	FP	2	519-211		
Yb8NBC17	AL049842	ACCATACAATTGGCAGAAC	GCAGGAAAGTGAATGTCGTGA	60	FP	2	507-184		
Yb8NBC18	AL078476	TGTGGTGTGGCTGGAGGATG	TGGGACTGATGTTCTGATGAA	56	IF	21	431-133		
Yb8NBC19	AC007077	TTGGGATGTAATTGTCGTGAG	AAACAGTCAGTGGGATCATG	55	FP	2	516-191		
Yb8NBC20	AC002059	TCTGGGCAAGTCACTAAA	ACTGTTTCCATGGGATCATG	56	FP	22	502-158		
Yb8NBC22	AC004875	CCACCAATGTCCCTTCTAC	AGTGGAAATCAGTCATTGGT	50	FP	7	598-275		
Yb8NBC23	AC005821	CCCTCCCTGGAAAGGCCAACAG	TGATCTTCAGGATGTTCC	60	FP	17	499-177		
Yb8NBC24	AC005772	AAACACTCCACCCATACTC	CATCTCTCTTCTCTGTTCC	56	FP	17	661-338		
Yb8NBC26	AP00036	GGGTCCTCTCTGGAGTAA	TTTTCCTGCCAACCAACAC	56	FP	21	499-150		
Yb8NBC27	AL021940	CTTGGGAGGGACACAGCTC	GGGGCACACAAATAATTCC	55	FP	1	522-188		
Yb8NBC28	AC003988	CTTTTGTATGAGTGTG	ACACAGAAACCCAGCTT	60	FP	5	515-197		
Yb8NBC29	AL09177	CTCTCGATGAGCAAGG	AGTGGCCAGAAACGAGATGTA	60	FP	6	519-200		
Yb8NBC30	Z95124	TTGCCCTGGATGGCATCT	AAATGGCGGAGTAAGTCTCT	55	IF	X	497-194		
Yb8NBC31	AC005046	TGAGGCAATTGGGGAGAG	TGGTGTGCTGTGTTTCCA	60	FP	7	518-188		
Yb8NBC32	AP000352	GGAGAGAAAGCACACATGT	GCAAAATGAGGAGGAAT	55	FP	22	496-167		
Yb8NBC34	AP000188	ATGACCTTTCGATTTACCA	GTTCAGCCACCCATCACATCT	60	FP	21	513-227		
Yb8NBC35	AP000171	CCTTTTGTCTCTGGCTGA	CCTGAAACACCCAGGAGAGA	60	FP	21	497-220		
Yb8NBC36	AP000156	CTAAAGGAGACGGGAGACAT	TAIATTCGATGACATGGCTTAC	60	FP	21	506-183		
Yb8NBC37	AP000193	AGAAATCGAGTTTCCGAGA	GGGATGGTCAAGATGGGATT	60	FP	21	495-162		
Yb8NBC38	AC002367	CGAGAGAAAGGGTAGAAAGC	AAATGCCATTCAAGGACATCTT	60	FP	X	486-311		
Yb8NBC39	AP000111	CATGTGTCATGTTGGTCAG	TAICCCCCACTTCGGATTAC	60	FP	21	496-129		
Yb8NBC40	AP000080	TOCCAGGAGGAGATGGAGAA	CCCTGTTAGGTTGCTGTTTACA	60	FP	8	500-171		
Yb8NBC41	AC006011	TGCTTAAGGTTGTCAGCAG	CAACCAGGAAATGCTGTTTACA	60	FP	8	482-153		
Yb8NBC42	AC005971	TTCCAGCATGTCAGAACATA	ATGGATCTTGAATGCTGTTGG	55	FP	17	517-189		
Yb8NBC43	AC004228	GTGGGACTGGCTCTACAG	GACAGCAGTTGCCATCTCA	60	FP	11	521-201		
Yb8NBC44	AC00521	ACACCCCTCTGGTACCAATT	ATGCCCTGCTGCTTACT	55	R/R	7	503-180		
Yb8NBC45	AB020874	TGGCTTCAAGAGCATATCCA	GAGAAATGAAACSCATTTGG	55	FP	9	567-274		
Yb8NBC46	AL031224	TGCAATTCTAAAGGAGTACCACTG	TCCATTCTGCTGCTCTCT	60	FP	6	523-203		
Yb8NBC48	AC005495	AGGGAAACGGTGTGTTGG	GAGGATGCAAAGGATGATGTGA	58	IF	17	511-178		
Yb8NBC49	AL031904	GCAGTGTGATTGGTTTCTG	GCTGAAAGAGGGATTGAAATC	60	IF	6	542-205		

Table 2. (continued)

Name	Accession	5' Primer sequence (5'-3')	3' Primer sequence (5'-3')	A:T*	Human diversity ^b	Chi: Loc. ^c	Product size ^b	Filled-empty
Y88NBC107	AC006222	GTTGGTTTCCGAGTGT	GACTGTTCACTGGTTGGAG	60	FP	4	527-207	
Y88NBC108	AF164343	TGTCACTTGATTGTCGATA	TCATGGCATCTGGAAACA	60	IF	Y	550-194	
Y88NBC109	AC006371	GTCGAACCTCACTGCTAAAT	CATGGTATTCGCAAGACTATGAC	55	IF	Y	532-212	
Y88NBC110	AC008383	AATGGCTGATCCGAACT	CTAGGTTCACTCCCTGCTT	60	LF	Y	507-186	
Y88NBC111	AC007320	CCAGTGTCACTCAGACTTTC	TACACACACACATGCTTCAAG	60	FP	Y	531-192	
Y88NBC112	AC006999	GCATCTTAACCTAACTCTGTC	CAGGACATAGGGTGTGAGTTACTA	60	FP	Y	503-192	
Y88NBC114	AC004617	GGGTGAGTAGTCACTGAAAGAGA	AGATCTTACCTTACCCCTTG	60	FP	Y	510-164	
Y88NBC115	AL010264	TCATTGCAACATGACCTGAC	GGCTGACITTTATCCACCTA	60	FP	22	28-297	
Y88NBC117	ZB2189	CAACCACACCCCTAAAGACAT	ACTCTGGTATGACCTGATAATAA	60	FP	22	461-111	
Y88NBC118	AC006548	GCAGAGACACATAAGCTGATTGAA	GTTCATGGCTTAAGCTGTAGTG	60	FP	22	519-200	
Y88NBC119	Z95114	AGACCTTTGTCAGATGGATAGTTG	GGAAAGGTTTCAAGGAAGAGTGS	60	IF	Y	425-110	
Y88NBC120	AC004019	ACGTCATCTCCATTACCTCTC	GGTGTAGAGCAAGAGACAGG	60	IF	22	532-212	
Y88NBC123	AL031846	TTTGGATTTGTCCTCT	AAAGGCCCAAGAAGTATACCA	60	IF	22	732-412	
Y88NBC125	DJ039122	AGCCAGAACCTCTGACAAAG	CCTAGCGACAAAGCCCTAGA	60	IF	22	415-97	
Y88NBC126	AC002055	AAATGTCCCCCCTGTCCTTC	CCCTGATTTCTTCAGCTG	60	IF	22	438-118	
Y88NBC129	AC004052	CCCAAACCTCTGATCTCC	TCATACAAAACCGAGCTGTT	55	FP	4	528-136	
Y88NBC131	AC002894	TGGGGTCACTTACTGACTCCA	AGCCAGCAAGAACCTGAGTC	60	FP	17	506-264	
Y88NBC132	AC002458	GTTCTGTGGTTGGATT	GGCTGTTAATGTCGCTCT	55	FP	4	507-187	
Y88NBC133	ZB4470	GCCATTGATCCACAGAAAT	AGGGTTCCAGTTCCCTATA	60	LF	X	536-232	
Y88NBC134	AC002067	TGACSCACAGATTGATAGGC	GGAAACAAAGGCAAAAGAGA	60	LF	7	526-206	
Y88NBC135	AC007392	TTTCCTCTCTCTGTTGAA	CTGATTTACCGGTTTCTCTC	60	FP	2	659-206	
Y88NBC136	AC007055	CTTGTCTCACCTGTTGGA	TCGAGCTATCTGGAGGAGT	55	FP	14	530-196	
Y88NBC137	AL031782	GGTGAAGTGGACGGCAA	TGCGCTTAACTGACCATCTT	60	FP	6	454-126	
Y88NBC138	AL031653	GTTCTCTCTCTCTCAAAG	CAGAAATGGTCTCTGTT	55	FP	20	650-332	
Y88NBC139	AC001612	TTTACTGGACAGTTGAAAC	TTTACTAGATGTCGAAAGGTT	60	RR	7	494-180	
Y88NBC140	AC003950	GGTAAGACAATAGTGGGATT	CATGGGTTCCCTCTGCTT	60	FP	17	530-219	
Y88NBC141	AL049869	TCAAGTGCCTCAGAAGCTG	CGTGTCTAAAGTGTGTTG	60	FP	14	487-162	
Y88NBC142	AC009044	GGCTCTTAAGCTAGACAATCAA	AGACGACTTATTGGGGATT	55	FP	16	443-133	
Y88NBC143	AL033531	TCAAGCTGTGCTTAACTCAA	CAGAACATGGCTGAGTGT	60	FP	1	578-255	
Y88NBC144	AL035089	TGGTCAGAACCTCTCAA	GGAGCTTAAACTGGCTCT	65	FP	20	520-197	
Y88NBC145	AC009028	CTCTCTCTCAGGAAACCTC	GTTGCTGTTGCTTACACTCAA	60	IF	16	887-592	
Y88NBC146	AC010340	GAATATGCTGCAATAGGAA	TGAGTCGTTAGATAGCAGTCAA	55	FP	5	516-149	
Y88NBC147	AC010582	CCAGCCCTCTCATTTGATA	TCATCTTGGCATGTCAGA	60	IF	14	537-218	
Y88NBC148	AL135746	TGAGTGAATGCAAAATCAAGG	TCAGAAACTCCGCAAGAGA	60	FP	14	460-138	
Y88NBC149	AF172277	CTGGCCATAAATCCCTCAG	TGTTGAAACTTACTCTGAGCT	60	FP	21	474-160	
Y88NBC150	AP00055	GGCAAGGAGGTCAAGTGA	ACTGGCAAGGGATGTGAGA	60	FP	21	464-138	
Y88NBC151	AP00056	ACTAGGGCCTTCACTCTGC	ATTCTCCCAAGGGAGTC	60	FP	18	520-183	
Y88NBC152	AL121248	TATGGTTCTCAACATAC	CAGGCTTAACTGGCTT	60	IF	10	712-423	
Y88NBC153	AC007571	GGAGGTTAACTGTCCTCAA	TGCTGTTGCTGTCTTAC	60	FP	12	569-248	
Y88NBC154	AC007880	GGCCTGAGCACTGTTAGTT	TATGCAAGGCTCCATGAGATA	60	FP	2	448-112	
Y88NBC155	AC007284	GGGAAGAGGTCAAGTGA	TTCCCTCTACTCCCTCATTC	60	IF	Y	404-72	
Y88NBC156	AC007100	ACATGGCAGCTTACAGTGTCA	CAAGCAAGTAAAGAAGTGC	60	FP	2	418-85	
Y88NBC157	AL132987	TATGGTTCTCAACATAC	TGCAAGTTCAGGCAATAACAG	60	FP	14	504-185	
Y88NBC158	AL035467	GCAGAGTAACTGCTGCT	TCACAGTCAAAAGCTTCAAG	60	FP	6	462-133	
Y88NBC159	AC009509	TGACACACATCCGTGACAGAA	TCACAGGCCAAATAAACAT	60	FP	12	387-76	
Y88NBC160	AC010200	TGGGTGAAGGGAGATTG	AACAGTGCCTAACTTCCTGAGAA	60	FP	12	465-151	

570-248	Y88N8C166	CTGCTGCCTCCCTAGACTG CCTCTGGCTCCAGGTTAA AGATGCCCACTATCAA TTGGTAGACTCCGGTAA TCCCCAAAGGAGGAGACA	55
AL121852	AL049777	TCACCTCTTAAGTGAACAGACTCAA ACTGGGTCTTCAAAAGTGG ATGGCATTCCTGGGTTCTTA GGCTCTAACCCAAATAGGA TTCCTCCATCCACAACTTTA	56
Y88N8C167	Y88N8C168	60	
Y88N8C168	Y88N8C169	60	
Y88N8C169	Y88N8C170	60	
Y88N8C170	AL109653	60	
Y88N8C171	AL096771	60	
Y88N8C172	AC00197	60	
Y88N8C173	AC007250	60	
Y88N8C173	AC011493	60	
Y88N8C177	AP000561	60	
Y88N8C178	AL080296	60	
Y88N8C181	AC007917	60	
Y88N8C183	AP000495	60	
Y88N8C184	AP008040	60	
Y88N8C185	AC008055	60	
Y88N8C186	AL031905	60	
Y88N8C187	AC007684	60	
Y88N8C189	AL078604	60	
Y88N8C191	AC006329	60	
Y88N8C192	AL049836	60	
Y88N8C193	AL109733	60	
Y88N8C195	AL036595	60	
Y88N8C198	AC007377	60	
Y88N8C199	AC008041	60	
Y88N8C200	AC007558	60	
Y88N8C201	AC006984	60	
Y88N8C202	AL031685	60	
Y88N8C203	AC004885	60	
Y88N8C204	AC007543	60	
Y88N8C205	AL033525	60	
Y88N8C206	AP000243	60	
Y88N8C208	AC006511	60	
Y88N8C209	AC007198	60	
Y88N8C210	AC007165	60	
Y88N8C211	AC006561	60	
Y88N8C212	AL033381	60	
Y88N8C213	AC001059	60	
Y88N8C214	AL034370	60	
Y88N8C216	AC005572	60	
Y88N8C217	AC005989	60	
Y88N8C217	AC005989	60	
Y88N8C218	AC004866	60	
Y88N8C219	AL024509	60	
Y88N8C220	AC005989	60	
Y88N8C221	AC004853	60	
Y88N8C222	AC005799	60	

Table 2. (continued)

Name	Accession	5' Primer sequence (5'-3')	3' Primer sequence (5'-3')	A.T.*	Human diversity ^b	Chr. ^c	Loc.	Product size ^b	Filled-empty
Y8NBBC28	AC005722	GTTGCCAGACCTGGCTCT	CCAGACAGCTGGGGTTT	62	FP	17	630-310		
Y8NBBC29	AC005754	CCAGAGTTCTACCTTGCACTG	TCACAACTGACACTCTTG	60	FP	5	411-90		
Y8NBBC30	AB014460	CAAAATGGCGTGTTCTT	GTGTCACAGGATCTTGAG	62	FP	16	458-124		
Y8NBBC31	AC005618	GGAAAGCTCCCTTCTTCAAG	ATGCAATTATTCCTCCACA	60	FP	5	501-181		
Y8NBBC32	AL023875	TGTGAACTCCACAGTCAGAA	TTCACAGCTGATCAGTCAA	55	FP	X	402-83		
Y8NBBC33	AC004702	TCCACATGGATGGAGAGATA	GTGGTCTGCAAGGAACTAGT	60	FP	17	446-114		
Y8NBBC34	AC005207	ACCTGAAAAGGGCGCTAGA	CTAATGGAGCCACCTCAA	60	FP	17	523-198		
Y8NBBC35	AC005221	CATTCTGGGACACTCCTACTT	CATCTAAATGCTTAAGGT	60	FP	5	473-146		
Y8NBBC36	AL021939	CTGCTTCACTGGCCAGAT	CAAAGCTTATGCTGCGTCA	60	FP	6	775-449		
Y8NBBC37	AC004613	GGCAAATCAACTGCCAAC	TGCTGAGGATAGAGCTATAGCAGA	60	FP	7	491-164		
Y8NBBC38	AC004592	AATGAGTCACCTGCCCTTG	CTCTGAGGATGGAGG	60	FP	5	437-117		
Y8NBBC39	AF031078	TTGCTGAGCATGAGGGATG	TCGCCCTCAAACTTATCC	55	FP	X	730-419		
Y8NBBC40	AC004652	TTCACAGTATGCTGCTICA	GGTGTCTCTGAGAATGCTA	60	FP	7	564-257		
Y8NBBC41	AC004391	GGACTGTGCTAAGGGTCTCT	GTAATGGGAGCAGTGTAGA	60	FP	7	424-93		
Y8NBBC42	AC002349	ATGGCTTCACTGGGACATT	TGAGCTGAGATAGCAGACATT	60	FP	X	450-117		
Y8NBBC44	AF043945	CGATGCTCCCTTACCAAA	CACTGCGTGGTCTACACTT	60	FP	21	493-77		
Y8NBBC45	AC004029	AACCCATGTCIOTATGCTAGC	CTCTCATCACAAAGTGTAGTGT	60	FP	7	647-318		
Y8NBBC46	AC002981	CACCAACCTTCAACCCAGAA	ATGCGTGTAAATGGTTCTC	60	FP	X	464-149		
Y8NBBC47	AC002366	GEAGCAAAAGTAGTGGTGG	GTAACCCACCTTATGATGCT	60	FP	X	551-259		
Y8NBBC48	AC003088	TTTCCTTCCCTGCGATGT	CCGCTTCTGCTGACACATT	60	FP	7	441-120		
Y8NBBC49	ZB8049	ATGGCCAAATAAAGGAT	GGTATGGCTCTGAGCAT	62	FP	6	491-148		
Y8NBBC50	AC002462	CGGATCCAGACATTGATT	TTGCTCTCTACTGCTCCCT	60	FP	?	397-63		
Y8NBBC51	AC002477	CGGCCCTGTATGCTCTTGA	TCCACAAAGGCTATGATA	60	FP	X	838-500		
Y8NBBC52	AC002123	GCCCCACCATCGAGTCATCT	TCCACATCTCCATCAGAGCTT	62	FP	5	424-107		
Y8NBBC53	AF01548	CAAAGGCAATCTGGAGCTG	CCCTCTCTCTCTTGTCA	60	FP	16	473-144		
Y8NBBC54	AC002088	GGGGAAACATTACTACAGGG	GGTATGGCTCTGAGTACGG	55	FP	6	740-413		
Y8NBBC55	AC000662	GGATGAAGTCTCCAGATGA	CGAGGAGGAGGAGCCAG	55	FP	13	423-93		
Y8NBBC56	Z73986	CCACACATTCTCCACTCTGG	GCATTGCTCTCTCCCTATTC	55	FP	X	503-24		
Y8NBBC57	Z69921	CTGCAACAAAAGAGACACAA	GCAAAACATAGAAAGGGGTGT	55	FP	4	508-187		
Y8NBBC58	AC004629	TTAGTGGTTCTGCGATGTT	AAGTTTATGCTCCCGCGCTGA	60	FP	-	431-103		
Y8NBBC59	AC015600	TCAGGAAAAGGGAGACATT	TTCAAGAAGAAAGGGAAACA	60	FP	?	547-227		
Y8NBBC60	AC012000	TTTCCACCATCAGCCCTCT	AGGGACTTGGATTGTTAGTGT	55	FP	2	661-327		
Y8NBBC61	AC009478	GCAGGACTTAATGCCAATCA	TCATCGTCTCTTACCTCTCTG	60	FP	2	375-50		
Y8NBBC62	AC020728	ATCCAGATTTCGAGGACCT	CCTTTGTCATCAACTACCACTG	60	FP	5	687-361		
Y8NBBC63	AC006318	GAAGAGGGAGGAGCATGT	AAGTTTATGCTCCCGCGCTGA	60	FP	12	518-177		
Y8NBBC64	AC007619	AGCAAGAACATATGCAAAAA	ATCGCTTCTTAAAGTGTGATACCA	60	FP	12	781-475		
Y8NBBC65	AC02436	TGCCAACTGACACTCTTG	CCCTTTGGATTCTCTGTC	60	FP	5	434-114		
Y8NBBC66	AC005078	TATTCATTCGCTCTTGG	ATGGCTCCACCTGTTAGTGG	60	FP	16	475-149		
Y8NBBC67	AC006925	GGGAAAGTTCAACAAACCAGA	CACCAACTGAAATGATCCCTTT	60	FP	5	777-447		
Y8NBBC68	AC016681	TGGGATAGAGGAGAGACAA	CCTTTGTCATCAACTACCACTG	60	FP	Y	517-188		
Y8NBBC69	AF241735	CACGCTTAACCTTACACCA	TGGACTCCACAGTGTG	60	FP	-	587-261		
Y8NBBC70	AC007489	CTTTCAGCTGCTGACCTG	AGTCTAGGCTTGGATGAG	55	FP	16	403-72		
Y8NBBC71	AC023602	GGAAAACCTGCTATGAGSC	CAGTTAAGTGTCTGCTG	60	FP	3	493-168		
Y8NBBC72	AC023602a	TGAGAAATGTTGCTTGTGG	TTCCCTGATCTGCTG	60	FP	-	537-215		
Y8NBBC73	AC011284	CTCCCTTGTGGGGAGAAG	CATGCTCTGGSGAACTCTC	55	FP	7	431-112		
Y8NBBC74	AL135305	TCACACATCAACCCACCTGA	TCCAGGGAGGAACTGAGATA	55	FP	14	489-168		
Y8NBBC75	AL122000	TTCTCTAAAGAGCTACACCTG	TTTGGGCTTATGAGGAAAG	60	FP	1	472-146		
Y8NBBC76	AL109928	TCACACATCAACCCACCTGA	GGCGGGTTGTTAAGTGGAAA	55	FP	20	507-183		
Y8NBBC77	AL121944	AAACATGAACACTGAAGGGAC	CCCTGCGAGCTGTATAATCA	60	FP	6	404-82		

* Amplification of each locus required 2:30 minutes at 94 °C initial denaturing, and 32 cycles for one minute at 94 °C, one minute at annealing temperature (A.T.), and one minute elongation at 72 °C. A final extension time of ten minutes at 72 °C was also used.

• Allele frequency was classified as: fixed present (FP), low (LF), intermediate (IF), or high frequency (HF) insertion polymorphism. Fixed present: every individual tested had the Alu element in both chromosomes. Low frequency insertion polymorphism: the absence of the element from all individuals tested, except for one or two homozygous or heterozygous individuals. Intermediate frequency insertion polymorphism: the Alu element is variable as to its presence or absence in at least one population. High frequency insertion polymorphism: the element is present in all individuals in the populations tested, except for one or two heterozygous or absent individuals. (-) Indeterminable. (R/R) Repeat in repeat.

^c Chromosomal location determined from Accession information or by PCR analysis of NIGMS monochromosomal hybrid cell line DNA samples.

^d Empty product sizes calculated by removing the Alu element and one direct repeat from the filled sites that were identified.

Table 3. Alu Ya5 subfamily associated human genomic diversity

Elements	African American				Greenland natives/Asian ^c				European				Egyptian													
	Genotypes				Genotypes				Genotypes				Genotypes													
	+/+	+-	-/-	fAlu	het ^a	++	+-	-/-	fAlu	het ^a	++	+-	-/-	fAlu	het ^a	+/+	+-	-/-	fAlu	het ^a	+/+	+-	-/-	fAlu	het ^a	Avg. het ^b
<i>A. Intermediate frequency</i>																										
Ya5NBC5	2	5	5	0.38	0.49	3	2	8	0.31	0.44	1	6	11	0.22	0.36	2	8	4	0.43	0.51	0.51	0.45	0.51	0.51	0.45	
Ya5NBC22	3	15	1	0.55	0.51	4	14	0	0.61	0.49	1	16	1	0.50	0.51	19	1	0	0.98	0.05	0.05	0.39	0.05	0.39	0.39	
Ya5NBC27	0	5	14	0.13	0.24	0	8	11	0.21	0.34	2	7	9	0.31	0.44	2	7	13	0	0.68	0.45	0.45	0.42	0.42	0.42	0.36
Ya5NBC35	9	10	1	0.70	0.43	5	12	2	0.58	0.50	8	12	0	0.70	0.43	0.33	4	3	10	0.32	0.45	0.45	0.34	0.45	0.45	0.34
Ya5NBC37	2	2	13	0.18	0.30	1	4	12	0.18	0.30	3	2	15	0.20	0.33	0	8	3	0	0.86	0.25	0.25	0.18	0.25	0.25	0.18
Ya5NBC45	7	7	2	0.66	0.47	19	0	0	1.00	0.00	17	0	0	1.00	0.00	8	3	0	0.86	0.25	0.25	0.18	0.25	0.25	0.18	
Ya5NBC51	4	10	3	0.53	0.51	5	6	8	0.42	0.50	6	7	7	0.48	0.51	3	8	9	0.35	0.47	0.47	0.50	0.47	0.47	0.50	
Ya5NBC57	10	1	2	0.81	0.32	4	8	3	0.53	0.52	13	2	1	0.88	0.23	9	1	1	0.86	0.25	0.25	0.33	0.25	0.25	0.33	
Ya5NBC61	10	6	3	0.68	0.44	5	2	10	0.35	0.47	9	7	1	0.74	0.40	8	4	5	0.59	0.50	0.50	0.45	0.50	0.50	0.45	
Ya5NBC96	17	2	0	0.95	0.10	9	5	3	0.68	0.45	18	1	0	0.97	0.05	16	3	0	0.92	0.15	0.15	0.19	0.15	0.15	0.19	
Ya5NBC102	3	2	13	0.22	0.36	0	0	6	0.00	0.00	3	4	12	0.26	0.40	2	0	13	0.13	0.24	0.24	0.25	0.24	0.24	0.25	
Ya5NBC109	7	11	1	0.66	0.46	7	11	2	0.63	0.48	5	13	1	0.61	0.49	7	8	4	0.58	0.48	0.48	0.48	0.48	0.48	0.48	
Ya5NBC120	7	11	0	0.69	0.44	15	4	0	0.90	0.19	8	12	0	0.70	0.43	14	5	0	0.87	0.24	0.24	0.32	0.24	0.24	0.32	
Ya5NBC123	5	7	7	0.45	0.51	6	5	4	0.57	0.51	14	5	1	0.83	0.30	11	5	1	0.79	0.34	0.34	0.41	0.34	0.34	0.41	
Ya5NBC131	0	5	6	0.23	0.37	0	9	8	0.27	0.40	0	11	6	0.32	0.45	0	15	2	0.44	0.51	0.51	0.43	0.51	0.51	0.43	
Ya5NBC132	4	0	5	0.44	0.52	9	0	0	1.00	0.00	13	0	0	1.00	0.00	11	0	1	0.92	0.05	0.05	0.159	0.05	0.05	0.159	
Ya5NBC148	7	6	6	0.53	0.51	2	6	12	0.25	0.39	0	0	20	0.00	0.00	0	0	0	0.00	0.00	0.00	0.22	0.00	0.00	0.22	
Ya5NBC150	17	0	0	1.00	0.00	4	0	14	0.22	0.36	19	0	1	0.95	0.00	17	0	1	0.94	0.11	0.11	0.14	0.11	0.11	0.14	
Ya5NBC154	0	12	5	0.35	0.47	0	7	9	0.22	0.35	0	12	8	0.30	0.43	3	4	13	0.25	0.39	0.39	0.41	0.39	0.39	0.41	
Ya5NBC160	2	7	9	0.31	0.44	0	0	19	0.00	0.00	0	0	19	0.00	0.00	0	4	12	0.13	0.23	0.23	0.17	0.23	0.23	0.17	
Ya5NBC174	5	3	31	0.46	0.46	0	3	8	0.14	0.77	0.25	0	12	8	0.30	0.43	2	5	9	0.28	0.42	0.42	0.39	0.42	0.42	0.39
Ya5NBC182	2	9	9	0.33	0.45	9	8	0	0.77	0.37	5	6	7	0.44	0.51	1	10	3	0.43	0.51	0.51	0.46	0.51	0.51	0.46	
Ya5NBC201	6	6	5	0.53	0.51	4	7	6	0.44	0.51	16	3	0	0.92	0.15	8	7	2	0.68	0.45	0.45	0.41	0.41	0.41	0.41	
Ya5NBC210	0	4	15	0.11	0.19	0	1	15	0.03	0.06	0	4	16	0.10	0.19	0	4	12	0.13	0.23	0.23	0.17	0.23	0.23	0.17	
Ya5NBC216	5	7	5	0.50	0.52	6	8	5	0.53	0.51	7	12	0	0.68	0.44	0	0	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Ya5NBC219	0	10	9	0.26	0.40	1	12	7	0.35	0.47	0	11	9	0.28	0.41	0	0	6	0.00	0.00	0.00	0.32	0.00	0.00	0.32	
Ya5NBC221	5	7	4	0.53	0.51	9	5	3	0.68	0.45	16	0	1	0.94	0.11	13	2	0	0.93	0.13	0.13	0.30	0.13	0.13	0.30	
Ya5NBC311 ^c	12	1	6	0.66	0.46	11	4	2	0.77	0.37	15	1	4	0.78	0.36	11	2	4	0.71	0.43	0.43	0.41	0.43	0.43	0.41	
Ya5NBC313 ^c	9	3	5	0.62	0.49	4	6	6	0.44	0.51	2	8	3	0.46	0.52	5	6	3	0.57	0.50	0.50	0.45	0.50	0.50	0.45	
Ya5NBC324 ^c	0	8	1	0.44	0.52	0	15	1	0.47	0.51	0	14	4	0.39	0.49	0	15	1	0.47	0.51	0.51	0.51	0.51	0.51	0.51	
Ya5NBC325 ^c	0	10	10	0.25	0.39	0	9	9	0.25	0.39	0	11	9	0.28	0.41	0	6	6	0.25	0.39	0.39	0.39	0.39	0.39	0.39	
Ya5NBC327 ^c	2	9	9	0.33	0.45	13	6	1	0.80	0.33	19	0	1	1.00	0.00	7	6	1	0.71	0.42	0.42	0.30	0.42	0.42	0.30	
Ya5NBC333 ^c	5	9	40	0.49	0.49	4	7	8	0.49	0.49	3	8	8	0.37	0.48	5	3	5	0.50	0.52	0.52	0.50	0.52	0.52	0.50	
Ya5NBC347 ^c	17	2	1	0.90	0.19	4	7	8	0.40	0.49	7	8	2	0.65	0.47	11	1	5	0.68	0.45	0.45	0.40	0.45	0.45	0.40	
Ya5NBC351 ^c	3	12	3	0.55	0.51	7	9	3	0.61	0.49	13	3	3	0.76	0.37	11	1	5	0.68	0.45	0.45	0.46	0.45	0.45	0.46	
Ya5NBC354 ^c	0	2	16	0.06	0.11	2	6	10	0.28	0.41	10	4	5	0.63	0.48	2	4	9	0.27	0.41	0.41	0.35	0.41	0.41	0.35	
Ya5NBC361 ^c	0	9	10	0.24	0.37	2	5	11	0.42	0.50	0	12	3	0.15	0.26	3	3	7	0.35	0.47	0.47	0.47	0.47	0.47	0.47	

B. <i>High frequency</i>	20	0	1.00	0.00	20	0	1.00	0.00	10	0	0.75	0.39	20	0	0	1.00	0.00	0	0	0.00	0.10	
Ya5NBC16	17	1	0	0.97	0.06	18	1	0	0.97	0.05	20	0	1.00	0.00	20	0	0	1.00	0.00	0	0	0.03
Ya5NBC18	17	0	1	0.94	0.11	18	2	0	0.95	0.10	17	1	2	0.88	0.22	16	3	0	0.92	0.15	0.14	
Ya5NBC98	17	0	1	1.00	0.00	20	0	0	1.00	0.00	20	0	0	1.00	0.00	15	0	1	0.94	0.12	0.03	
Ya5NBC157	20	0	0	1.00	0.00	20	0	0	1.00	0.00	20	0	0	1.00	0.00	18	0	0	1.00	0.00	0.01	
Ya5NBC212	16	1	0	0.97	0.06	20	0	0	1.00	0.00	20	0	0	1.00	0.00	12	0	2	0.86	0.25	0.17	
Ya5NBC349 ^c	19	1	0	0.98	0.05	14	2	4	0.75	0.39	18	0	0	1.00	0.00	12	0	2	0.86	0.25	0.17	
C. <i>Low frequency</i>																						
Ya5NBC24	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	0.00	0.00	
Ya5NBC28	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	0.00	0.00	
Ya5NBC38	0	0	16	0.00	0.00	0	0	15	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	0.00	0.00	
Ya5NBC54	0	0	14	0.00	0.00	0	0	2	0.07	0.14	0	0	1	0.06	0.11	0	0	1	0.06	0.13	0.09	
Ya5NBC135	0	1	18	0.03	0.05	0	0	20	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	0.00	0.01	
Ya5NBC147	0	1	17	0.03	0.05	0	1	19	0.03	0.05	0	0	0	0.00	0	20	0.00	0	1	0.03	0.06	
Ya5NBC155	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	0.00	0.00	
Ya5NBC171	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	0.00	0.00	
Ya5NBC172	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	0.00	0.00	
Ya5NBC184	0	0	20	0.00	0.00	0	0	18	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	0.00	0.00	
Ya5NBC194	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	0.00	0.00	
Ya5NBC197	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	0.00	0.00	
Ya5NBC203	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	0.00	0.00	
Ya5NBC204	0	0	19	0.00	0.00	0	0	17	0.00	0.00	0	0	0	0.00	0	17	0.00	0	0	0.00	0.00	
Ya5NBC214	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	5	15	0.13	0.22	0	0	19	0.00	0.00	0.06	
Ya5NBC223	0	0	20	0.00	0.00	0	0	20	0.00	0.00	0	0	0	0.00	0	20	0.00	0	0	19	0.00	

^a This is the unbiased heterozygosity.^b Average heterozygosity is the average of the population heterozygosity.^c The following were tested using DNA samples from Asian individuals.

Table 4. Alu Yb8 subfamily associated human genomic diversity

Elements	African American				Greenland natives/Asian ^a				European				Egyptian							
	Genotypes		Genotypes		Genotypes		Genotypes		Genotypes		Genotypes		Genotypes		Genotypes		Genotypes			
	+/+	+/−	−/−	fAlu	het ^b	+/+	+/−	−/−	fAlu	het ^b	+/+	+/−	−/−	fAlu	het ^b	+/−	−/−	fAlu	het ^b	
A. Intermediate frequency																				
Yb8NBC3	10	2	0	0.92	0.16	12	4	0	0.88	0.23	4	6	0	0.70	0.44	9	0	1.00	0.00	0.207
Yb8NBC7	5	8	0	0.69	0.51	4	14	0	0.61	0.49	1	16	1	0.50	0.51	19	1	0	0.98	0.05
Yb8NBC8	8	1	0	0.94	0.11	10	4	0	0.86	0.25	11	1	2	0.82	0.30	12	2	3	0.77	0.26
Yb8NBC9	3	5	10	0.31	0.44	2	3	13	0.19	0.32	5	1	9	0.37	0.48	0	7	8	0.23	0.37
Yb8NBC10	9	9	0	0.75	0.39	9	11	0	0.73	0.41	12	7	0	0.82	0.31	11	5	0	0.84	0.27
Yb8NBC18	1	0	15	0.06	0.12	2	9	9	0.33	0.45	0	6	14	0.15	0.26	1	6	11	0.22	0.05
Yb8NBC30	8	6	0	0.79	0.35	7	11	0	0.69	0.44	5	8	0	0.69	0.44	14	2	0	0.94	0.12
Yb8NBC36	5	14	1	0.60	0.49	8	0	1.00	0.00	10	9	0	0.76	0.37	8	8	0	0.75	0.39	
Yb8NBC48	0	4	6	0.20	0.34	0	1	2	0.17	0.33	0	3	4	0.21	0.36	0	2	3	0.20	0.36
Yb8NBC49	1	9	10	0.28	0.41	7	8	5	0.55	0.51	5	9	6	0.48	0.51	1	8	9	0.28	0.41
Yb8NBC65	7	6	5	0.56	0.51	3	10	7	0.40	0.49	7	4	9	0.45	0.51	2	5	9	0.42	0.481
Yb8NBC67	8	5	5	0.58	0.50	9	6	4	0.63	0.48	13	2	0	0.93	0.13	4	7	4	0.50	0.52
Yb8NBC71	0	3	13	0.09	0.18	3	10	0	0.28	0.42	0	5	12	0.15	0.26	2	2	9	0.23	0.37
Yb8NBC77	2	2	16	0.15	0.26	2	0	16	0.11	0.20	0	1	17	0.03	0.06	0	0	16	0.00	0.13
Yb8NBC80	1	4	15	0.15	0.26	2	5	12	0.24	0.37	3	1	15	0.18	0.31	2	5	8	0.30	0.43
Yb8NBC93	1	3	10	0.18	0.30	7	5	2	0.18	0.30	7	2	5	0.57	0.51	12	4	1	0.82	0.35
Yb8NBC96	0	7	9	0.22	0.35	0	14	3	0.41	0.50	0	3	15	0.08	0.16	0	5	7	0.21	0.34
Yb8NBC106	4	6	7	0.41	0.50	2	8	10	0.30	0.43	0	2	18	0.05	0.10	3	5	11	0.29	0.42
Yb8NBC108	2	11	7	0.38	0.48	2	10	7	0.37	0.48	0	3	11	0.11	0.20	3	4	10	0.29	0.43
Yb8NBC109	0	11	8	0.29	0.42	1	11	8	0.33	0.45	4	1	6	0.41	0.51	7	0	11	0.39	0.467
Yb8NBC120	5	8	5	0.50	0.51	5	6	8	0.42	0.50	8	7	3	0.64	0.48	4	2	6	0.42	0.51
Yb8NBC125	0	0	20	0.00	0.00	0	3	16	0.08	0.15	0	3	17	0.08	0.14	0	5	14	0.13	0.24
Yb8NBC146	18	0	2	0.90	0.19	12	1	1	0.89	0.20	16	0	2	0.89	0.20	10	1	6	0.62	0.49
Yb8NBC148	11	0	2	0.85	0.27	11	1	6	0.64	0.48	6	2	10	0.39	0.49	13	3	4	0.75	0.411
Yb8NBC157	19	0	1	0.95	0.05	6	5	1	0.71	0.43	3	9	6	0.42	0.50	16	2	2	0.85	0.22
Yb8NBC160	0	12	8	0.25	0.39	0	13	7	0.33	0.45	0	10	10	0.25	0.39	1	6	13	0.20	0.33
Yb8NBC189	10	10	0	0.75	0.39	18	2	0	0.95	0.10	9	7	2	0.69	0.44	18	2	0	0.95	0.10
Yb8NBC201	5	9	0	0.59	0.50	3	8	7	0.39	0.49	9	5	6	0.58	0.50	2	9	8	0.34	0.46
Yb8NBC208	5	6	5	0.50	0.52	18	2	1	0.91	0.18	10	8	2	0.70	0.43	15	4	1	0.85	0.26
Yb8NBC225	10	9	1	0.73	0.41	12	2	4	0.72	0.41	11	6	3	0.70	0.43	8	2	5	0.60	0.50
Yb8NBC227 ^c	10	8	2	0.70	0.43	5	6	5	0.50	0.52	18	2	0	0.95	0.10	15	4	1	0.85	0.26
Yb8NBC230 ^c	1	2	11	0.14	0.25	0	0	19	0.00	0.00	0	2	15	0.06	0.11	1	4	3	0.38	0.50
Yb8NBC237 ^c	13	4	1	0.83	0.29	12	5	2	0.76	0.37	15	2	0	0.94	0.11	10	8	1	0.74	0.40
Yb8NBC241 ^c	0	16	0	0.00	0.00	2	0	14	0.13	0.23	2	3	10	0.23	0.37	1	6	8	0.27	0.41
Yb8NBC268 ^c	0	13	5	0.36	0.48	0	16	7	0.18	0.18	12	8	8	0.31	0.44	0	5	12	0.15	0.26

3. High frequency		4. Low frequency	
Y58NBC24	2	0	0.93
Y58NBC26	12	1	0.96
Y58NBC102	16	0	0.89
Y58NBC181	14	0	1.00
Y58NBC192	20	0	1.00
Y58NBC5	3	4	0.47
Y58NBC13	15	3	0
Y58NBC69	0	0	18
Y58NBC100	0	1	17
Y58NBC110	0	0	18
Y58NBC126	1	6	7
Y58NBC133	0	2	18
Y58NBC134	0	4	18

- This is the unbiased heterozygosity:

- b Average heterozygosity is the average of the population heterozygosity.
- c The following were tested using DNA samples from Asian individuals.

- ↳ Average heterozygosity is the average of the population heterozygosity.

ancestral state.^{35,36} Previously, the analysis of Alu insertion polymorphisms has proved useful for the study of human population genetics.³⁵⁻⁴³ The newly identified Alu insertion polymorphisms from the Ya5 and Yb8 Alu subfamilies should prove useful for the study of human population genetics.

Materials and Methods

Cell lines and DNA samples

The cell lines used to isolate primate DNA samples were as follows: human (*Homo sapiens*), HeLa (ATCC CCL2); and chimpanzee (*Pan troglodytes*), Wes (ATCC CRL1609). Cell lines were maintained as directed by the source and DNA isolations were performed using Wizard genomic DNA purification (Promega). Human DNA samples from the European, African American, Asian, Egyptian, and Greenland Native population groups were isolated from peripheral blood lymphocytes⁴⁴ available from previous studies.¹⁸

Computational analyses

Initial screening of the GenBank non-redundant and high throughput genomic sequence (HTGS) databases was performed using the Basic Local Alignment Search Tool (BLAST)⁴⁵ available from the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). Copy number estimates were determined using Megablast and the draft human genome sequence database.⁴⁶ The database was searched for exact complements to the oligonucleotide 5'-CCATCCC-GGCTAAAC-3' and 5'-TGCAGCCACTGCAGTCG-CAGTCGG-3' that are exact matches to a portion of the Alu Ya5 and Yb8 subfamily consensus sequences (respectively) that contain unique diagnostic mutations.²¹ Sequences that were exact complements to the oligonucleotides were then subjected to more detailed annotation. A region composed of 500-1000 bases of flanking DNA sequence directly adjacent to the sequences identified from the databases that matched the initial GenBank BLAST query were subjected to annotation using the RepeatMasker2 program from the University of Washington Genome Center server (<http://ftp.genome.washington.edu/c/s.dll/RepeatMasker>) or Censor from the Genetic Information Research Institute (http://www.girinst.org/Censor_Server-Data_Entry_Forms.html).⁴⁷ These programs annotate the repeat sequence content of individual sequences from humans and rodents. A complete list of the Alu elements identified from the GenBank search is available from MAB. The copy numbers for each subfamily of Alu elements were determined by screening the draft sequence of the entire human genome with the oligonucleotides shown above.²³ For the Yb8 subfamily analysis, the database was searched for matches to the consensus Yb8 sequence without the seven-nucleotide duplication (287 bases). The sequences were then subjected to more detailed analysis using MegAlign (DNASTar version 3.1.7 for Windows 3.2) selecting only for Yb8 intermediate elements containing between one and seven of the Yb8 diagnostic sites.

Primer design and PCR amplification

PCR primers were designed from flanking unique DNA sequences adjacent to individual Ya5 and Yb8 Alu elements using the Primer3 software (Whitehead Institute for Biomedical Research, Cambridge, MA, USA) (http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). The resultant PCR primers were screened against the GenBank non-redundant database for the presence of repetitive elements using the BLAST program, and primers that resided within known repetitive elements were discarded and new primers were designed. PCR amplification was carried out in 25 μ l reactions using 50-100 ng of target DNA, 40 pM of each oligonucleotide primer, 200 μ M dNTPs in 50 mM KCl, 1.5 mM MgCl₂, 10 mM Tris-HCl (pH 8.4) and *Taq*[®] DNA polymerase (1.25 units) as recommended by the supplier (Life Technologies). Each sample was subjected to the following amplification cycle: an initial denaturation of 150 seconds at 94°C, one minute of denaturation at 94°C, one minute at the annealing temperature, one minute of extension at 72°C, repeated for 32 cycles, followed by a final extension at 72°C for ten minutes. For analysis, 20 μ l of each sample was fractionated on a 2% agarose gel with 0.25 μ g/ml ethidium bromide. PCR products were directly visualized using UV fluorescence. The sequences of the oligonucleotide primers, annealing temperatures, PCR product sizes and chromosomal locations for all Ya5 and Yb8 elements can be found on our website (<http://129.81.225.52>). Phylogenetic analysis of all the ascertained Alu elements was determined by PCR amplification of human and non-human primate DNA samples. The human genomic diversity associated with each Alu element was determined by the amplification of 20 individuals from each of four populations (African-American, Greenland Native or Asian, European and Egyptian) (160 total chromosomes). The chromosomal location of Alu repeats identified from clones that had not been previously mapped was determined by PCR amplification of National Institute of General Medical Sciences (NIGMS) human/rodent somatic cell hybrid mapping panel 2 (Coriell Institute for Medical Research, Camden, NJ).

Acknowledgments

A.M.R. was supported by a Brown Foundation fellowship from the Tulane Cancer Center. This research was supported by National Institutes of Health RO1 GM45668 (P.L.D.) and RO1 GM59290 (L.B.J. and M.A.B.), Department of the Army DAMD17-98-1-8119 (P.L.D. and M.A.B.), and Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05)-05 and HEF (2000-05)-01 (M.A.B. and P.L.D.).

References

1. Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* 9, 657-663.
2. Deininger, P. L. & Batzer, M. A. (1993). Evolution of retrotransposons. In *Evolutionary Biology* (Heckht, M. K., et al. ed.), pp. 157-196, Plenum Publishing, New York.
3. Maklalowski, W., Mitchell, G. & Labuda, D. (1994). Alu sequences in the coding regions of mRNA: a

and human full-length mRNA and protein sequences. *Genome Res.* **6**, 846-857.

5. Weiner, A., Deininger, P. L. & Efstratiadis, A. (1986). The reverse flow of genetic information: pseudogenes and transposable elements derived from non-viral cellular RNA. *Annu. Rev. Biochem.* **55**, 631-661.
6. Boeke, J. (1997). LINEs and Alus - the polyA connection. *Nature Genet.* **16**, 37-43.
7. Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci. USA*, **94**, 1872-1877.
8. Sinnott, D., Richer, C., Deragon, J. M. & Labuda, D. (1992). Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences. *J. Mol. Biol.* **226**, 689-706.
9. Batzer, M. A., Kilroy, G. E., Richard, P. E., Shaikh, T. H., Dessel, T. D. et al. (1990). Structure and variability of recently inserted Alu family members. *Nucl. Acids Res.* **18**, 6793-6798.
10. Deininger, P. L., Batzer, M. A., Hutchison, C. & Edgell, M. (1992). Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**, 307-312.
11. Deininger, P. L. & Daniels, G. (1986). The recent evolution of mammalian repetitive DNA elements. *Trends Genet.* **2**, 76-80.
12. Leeflang, E. P., Liu, W-M., Hashimoto, C., Choudary, P. V. & Schmid, C. W. (1992). Phylogenetic evidence for multiple Alu source genes. *J. Mol. Evol.* **35**, 7-16.
13. Leeflang, E. P., Liu, W-M., Chesnokov, I. N. & Schmid, C. W. (1993). Phylogenetic isolation of a human Alu flounder gene: drift to new subfamily identity. *J. Mol. Evol.* **37**, 559-565.
14. Schmid, C. & Maraia, R. (1992). Transcriptional and transpositional selection of active SINE sequences. *Curr. Opin. Genet. Dev.* **2**, 874-882.
15. Shen, M., Batzer, M. A. & Deininger, P. L. (1991). Evolution of the Master Alu Gene(s). *J. Mol. Evol.* **33**, 311-320.
16. Batzer, M. A., Deininger, P. L., Hellmann-Blumberg, U., Jurka, J., Labuda, D. et al. (1996). Standardized nomenclature for Alu repeats. *J. Mol. Evol.* **42**, 3-6.
17. Roy, A. M., Carroll, M. L., Kass, D. H., Nguyen, S. V., Salem, A-H. et al. (1999). Recently integrated human Alu repeats: finding needles in the haystack. *Genetica*, **107**, 1-13.
18. Roy, A. M., Carroll, M. L., Nguyen, S. V., Salem, A-H., Oldridge, M. et al. (2000). Potential gene conversion and source gene(s) for recently integrated Alu elements. *Genome Res.* **10**, 1485-1495.
19. Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H., Jr. & Deininger, P. L. (1987). Clustering and subfamily relationships of the Alu family in the human genome. *Mol. Biol. Evol.* **4**, 19-29.
20. Willard, C., Nguyen, H. T. & Schmid, C. W. (1987). Existence of at least three distinct Alu subfamilies. *J. Mol. Evol.* **26**, 180-186.
21. Batzer, M. A., Rubin, C. M., Hellmann-Blumberg, U., Alegria-Hartman, M., Leeflang, E. P. et al. (1995). Dispersion and insertion polymorphism in two small subfamilies of recently amplified Alu repeats. *J. Mol. Biol.* **247**, 418-427.
22. Deininger, P. L. & Batzer, M. A. (1999). Alu Repeats and human disease. *Mol. Genet. Metab.* **67**, 183-193.
23. Lander, E. S., Linton, L. M., Birren, B. & Nusbaum, R. et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.
24. Arcot, S. S., Shaikh, T. H., Kim, J., Bennett, L., Alegria-Hartman, M. et al. (1995). Sequence diversity and chromosomal distribution of "young" Alu repeats. *Gene*, **163**, 273-278.
25. Labuda, D. & Striker, G. (1989). Sequence conservation in Alu evolution. *Nucl. Acids Res.* **17**, 2477-2491.
26. Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucl. Acids Res.* **8**, 1499-1504.
27. Miyamoto, M. M., Slightom, J. L. & Goodman, M. (1987). Phylogenetic relations of humans and African apes from DNA sequences in the pseudo- η -globin region. *Science*, **238**, 369-373.
28. Arcot, S. S., Wang, Z., Weber, J. L., Deininger, P. L. & Batzer, M. A. (1995). Alu repeats: a source for the genesis of primate microsatellites. *Genomics*, **29**, 136-144.
29. Maeda, N., Wu, C.-I., Bliska, J. & Renneke, J. (1988). Molecular evolution of intergenic DNA in higher primates: pattern of DNA changes, molecular clock, and evolution of repetitive sequences. *Mol. Biol. Evol.* **5**, 1-20.
30. Kass, D. H., Batzer, M. A. & Deininger, P. L. (1995). Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. *Mol. Cell. Biol.* **15**, 19-25.
31. Batzer, M. A., Gudi, V., Mena, J. C., Foltz, D. W., Herrera, R. J. & Deininger, P. L. (1991). Amplification dynamics of human-specific (HS) Alu family members. *Nucl. Acids Res.* **19**, 3619-3623.
32. Hutchinson, G. B., Andrew, S. E., McDonald, H., Goldberg, Y. P., Grahm, R. et al. (1993). An Alu element retrotransposition in two families with Huntington disease defines a new active Alu subfamily. *Nucl. Acids Res.* **21**, 3379-3383.
33. Edwards, M. C. & Gibbs, R. A. (1992). A human dimorphism resulting from loss of an Alu. *Genomics*, **3**, 590-597.
34. Sheen, F.-M., Sherry, S. T., Risch, G. M., Robichaux, M., Nasidze, I. et al. (2000). Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. *Genome Res.* **10**, 1496-1508.
35. Batzer, M. A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D. H. et al. (1994). African origin of human-specific polymorphic Alu insertions. *Proc. Natl. Acad. Sci. USA*, **91**, 12288-12292.
36. Stoneking, M., Fontius, J. J., Clifford, S. L., Soodyall, H., Arcot, S. S. et al. (1997). Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* **7**, 1061-1071.
37. Comas, D., Calafell, F., Benchemsi, N., Helal, A. & Lefranc, G. et al. (2000). Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. *Hum. Genet.* **107**, 312-319.
38. Hammer, M. F. (1994). A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. *Mol. Biol. Evol.* **11**, 749-761.
39. Jorde, L. B., Watkins, W. S., Bamshad, M. J., Dixon, M. E., Ricker, C. E. et al. (2000). The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am. J. Hum. Genet.* **66**, 979-988.
40. Majumder, P. P., Roy, B., Banerjee, S., Chakraborty, M., Dey, B. et al. (1999). Human-specific insertion/deletion polymorphisms in Indian populations and

their possible evolutionary implications. *Eur. J. Hum. Genet.* **7**, 435-446.

41. Perna, N. T., Batzer, M. A., Deininger, P. L. & Stoneking, M. (1992). Alu insertion polymorphism: a new type of marker for human population studies. *Hum. Biol.* **64**, 641-648.
42. Tishkoff, S. A., Ruano, G., Kidd, J. R. & Kidd, K. K. (1996). Distribution and frequency of a polymorphic Alu insertion at the plasminogen activator locus in humans. *Hum. Genet.* **97**, 759-764.
43. Watkins, W. S., Ricker, C. E., Bamshad, M. J., Carroll, M. L., Nguyen, S. V. et al. (2001). Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. *Am. J. Hum. Genet.* **68**, 738-752.
44. Ausabel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G. et al. (1987). In *Current Protocols in Molecular Biology*, New York.
45. Altschul, S., Madden, T., Schaffer, A., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
46. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203-214.
47. Jurka, J. P., Klonowski, Dagman, V. & Pelton, P. (1996). CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20**, 119-121.

Edited by J. Karm

(Received 9 March 2001; received in revised form 7 July 2001; accepted 11 July 2001)

Alu-Insertion Polymorphisms for the Study of Human Genomic Diversity

Astrid M. Roy-Engel,^{*,†} Marion L. Carroll,^{†,‡} Erika Vogel,^{*} Randall K. Garber,^{†,‡} Son V. Nguyen,[†] Abdel-Halim Salem,^{†,‡} Mark A. Batzer^{†,‡,§} and Prescott L. Deininger^{*,§,‡}

^{*}Tulane Cancer Center, Department of Environmental Health Sciences, Tulane University Health Sciences Center, New Orleans, Louisiana 70112, [†]Departments of Pathology, Genetics, Biochemistry and Molecular Biology, Stanley S. Scott Cancer Center, Louisiana State University Health Sciences Center, New Orleans, Louisiana 70112, [‡]Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, Baton Rouge, Louisiana 70803 and [§]Laboratory of Molecular Genetics, Alton Ochsner Medical Foundation, New Orleans, Louisiana 70121

Manuscript received February 24, 2001
Accepted for publication June 8, 2001

ABSTRACT

Genomic database mining has been a very useful aid in the identification and retrieval of recently integrated Alu elements from the human genome. We analyzed Alu elements retrieved from the GenBank database and identified two new Alu subfamilies, Alu Yb9 and Alu Yc2, and further characterized Yc1 subfamily members. Some members of each of the three subfamilies have inserted in the human genome so recently that about a one-third of the analyzed elements are polymorphic for the presence/absence of the Alu repeat in diverse human populations. These newly identified Alu insertion polymorphisms will serve as identical-by-descent genetic markers for the study of human evolution and forensics. Three previously classified Alu Y elements linked with disease belong to the Yc1 subfamily, supporting the retroposition potential of this subfamily and demonstrating that the Alu Y subfamily currently has a very low amplification rate in the human genome.

Alu elements have been accumulating in the human genome throughout primate evolution, reaching a copy number of over a million per genome. However, most of these Alu copies are not identical and can be classified into several subfamilies (reviewed in DEININGER and BATZER 1993). These different subfamilies of Alu elements were generated once mutations occurred within the "master" or "source" gene that actively retroposed at different rates and time periods of primate evolution (DEININGER *et al.* 1992). Currently, the Alu retroposition rate is reduced by 100-fold from its peak early in primate evolution (SHEN *et al.* 1991). The vast majority of the Alu elements present in the human genome inserted before the radiation of extant humans and are therefore observed in all individuals in the human population. However, almost all of the recently integrated Alu elements in the human genome are restricted to several closely related "young" subfamilies, with the majority being Ya5 and Yb8 subfamily members (BATZER *et al.* 1994, 1995). Several of these new subfamilies appear to originate from an Alu element that fortuitously inserted into a favorable region of the genome capable of supporting Alu retroposition. Subsequent or concurrent mutations in the new source element(s)

result in groups of elements that are identifiable as new subfamilies.

Collectively, the Alu Y, Ya5, Ya5a2, Ya8, and Yb8 subfamilies comprise <10% of the Alu elements present within the human genome, with the Ya5/8 and Yb8 subfamilies together accounting for <0.5% of all Alu elements. Although the human genome contains >1,000,000 copies of Alu (~15% of the genome; SMIT 1996), <0.5% are polymorphic. Due to their recent evolutionary introduction into the human genome, many of the young Alu elements are polymorphic between individuals and/or populations. There is an inverse correlation between the age of the Alu subfamily and the percentage of polymorphic elements it contains. Identification of evolutionarily recent Alu subfamilies and their polymorphic insertions is useful for human population studies, forensics, and DNA fingerprinting for two reasons: (i) There is no apparent specific mechanism to remove newly inserted Alu repeats, making inserts identical by descent; and (ii) the Alu insertions have a known ancestral state (BATZER and DEININGER 1991; BATZER *et al.* 1994).

The availability of large quantities of human genomic DNA sequence provided by the Human Genome Project facilitates genomic database mining for recently integrated Alu elements. Through this approach we were able to identify the youngest Alu subfamily reported to date, termed (Ya5a2), and determined that the majority of its members are Alu insertion polymorphisms (ROY *et al.* 2000). We expanded our computational analyses to identify other Alu subfamilies derived from the Alu

Corresponding author: Prescott L. Deininger, Tulane Cancer Center, Tulane University Medical Center, 1430 Tulane Ave., SL-66, New Orleans, LA 70112. E-mail: pdeinin@tcs.tulane.edu

[†]These authors contributed equally to this work.

[‡]These are equal senior authors.

Y and Yb8 subfamilies. Here, we present the analysis of three of the most recently formed Alu subfamilies and demonstrate their utility for the study of human genomic diversity.

MATERIALS AND METHODS

Computational analyses: Sequence alignments for the identification of Alu subfamilies were made using MegAlign software (DNAStar version 3.1.7 for Windows 3.2). Screening of the GenBank nonredundant (nr), the high throughput genome sequence (htgs), and the genomic survey sequence (gss) databases was performed using the advanced basic local alignment search tool 2.0 (BLAST; ALTSCHUL *et al.* 1990) available from the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). Database searches for Yb8 consensus Alus showed a common single-base variant termed Yb9. The databases were searched for matches to the 289 bases of the Yb9 consensus sequence (as inferred from the previous Yb8 analysis) or the 281 bases of the Alu Y consensus with the expected value (real) set at $-e 1.0e^{-150}$ and $-e 1.0e^{-140}$, respectively, in the advanced BLAST options. Only Alu Yb9 elements with all nine diagnostic mutations were selected. A similar type of search procedure was performed with the Yc1 and Yc2 consensus sequences or with an oligonucleotide query sequence complementary to the subfamily diagnostic base positions. Only Alu Yc1/Yc2 elements with 100% identity to the oligonucleotide query sequences or entire subfamily-specific consensus sequence were utilized for further analysis. To estimate the copy numbers of the Yb9 subfamily we searched the draft sequence of the human genome (LANDER *et al.* 2001), using a subfamily-specific probe that contained the Yb9-specific mutation as well as the insertion in the Yb8 subfamily. A complete list of the Alu elements identified from the GenBank search is available from M. A. Batzer or P. L. Deininger.

DNA samples: Human DNA samples from the European, African-American, Alaskan Native, Egyptian, and Asian population groups were isolated from peripheral blood lymphocytes (AUSUBEL *et al.* 1996) that were available from previous studies (ROY *et al.* 1999).

Oligonucleotide primer design and PCR amplification: Flanking unique DNA sequences adjacent to each Alu repeat were used to design primers for the Yb9, Yc1, and Yc2 Alu elements (Table 1). PCR primers and reactions were performed as previously described (ROY *et al.* 1999). The heterozygosity associated with each element was determined by the amplification of 20 individuals from each of four populations (African American, Alaskan Native, or Asian, European, and Egyptian; 160 total chromosomes). The chromosomal location for elements identified from randomly sequenced anonymous large-insert clones was determined by PCR as previously described (ROY *et al.* 1999).

RESULTS

The Alu Yb9, Yc1, and Yc2 subfamilies: Analysis of a set of 243 Yb8 Alu elements retrieved from the GenBank database allowed us to identify a putative subfamily containing all the known Yb8 diagnostic mutations plus one new mutation, which is referred to as Yb9 in compliance with the standard Alu subfamily nomenclature (BATZER *et al.* 1996). The Yb9 consensus sequence is shown in Figure 1. Searches from the nr, the htgs, and gss retrieved a total of 56 Yb9 elements. Of these, 25 elements

were retrieved from the nr database (30.4% of the human genome at the time), giving an estimated size of 82 members for the Yb9 subfamily. This estimate is also in good agreement with a search of the draft human genomic sequence (LANDER *et al.* 2001) that identified 79 perfect matches with a Yb9 subfamily-specific query sequence.

Using a different approach, we also retrieved one previously identified subfamily, Yc1 [formerly termed Sb0 (JURKA 1995)], and a new variant, Yc2. GenBank database searches for Alu Y elements that perfectly match the consensus sequence brought several Alu Y elements to our attention that share one or two specific mutations that differ from the Y consensus. Closer inspection facilitated the retrieval of the additional Alu subfamilies. BLAST searches using the consensus sequence for Alu Yc1 and Yc2 will also retrieve a large number of elements that are matches to the Alu Y subfamily as well, making the analysis of the elements identified in this manner impractical. Therefore, we selected only the elements of these subfamilies with 100% identity to the oligonucleotide query sequence that contained the subfamily-specific diagnostic bases. A total of 176 Yc1 (13 perfect matches to the entire subfamily consensus sequence) and 17 Yc2 (11 perfect matches to the entire subfamily consensus sequence) elements were retrieved. A count of all Yc1 elements retrieved by BLAST on a single initial search of the nr database yielded a total of 116 elements, giving an estimated copy number of 381 Yc1 elements in the human genome (the nr database contained 30.4% of the human genome sequence at the time of the search). Interestingly, three of the four elements previously classified as Alu Y elements linked to disease (DEININGER and BATZER 1999) belong to the Alu Yc1 subfamily (Figure 2): the *de novo* insertion in the C1 inhibitor gene (Clinh; STOPPAPYONNET *et al.* 1990), another *de novo* insertion in BRCA2 (BRCA2; MIKI *et al.* 1996), and glycerol kinase deficiency (GK; ZHANG *et al.* 2000).

About one-half of the 56 total Yb9 elements (29) shared 100% nucleotide identity with the subfamily consensus sequence. To get an approximation of the age of the Yb9 subfamily, we evaluated the number of non-CpG mutations present within the different Alu elements as previously described (ROY *et al.* 2000). A total of 19 CpG mutations, 25 non-CpG mutations, and two 5' truncations occurred within the 56 Alu Yb9 subfamily members identified. Using a neutral rate of evolution for primate intervening DNA sequences of 0.15% per million years (MIYAMOTO *et al.* 1987) and the non-CpG mutation density of 0.1908% (25/13,104 bases using only non-CpG bases) within the 56 Yb9 Alu elements yield an estimated average age of 1.27 million years (myr). The age for the Yb9 subfamily members is predicted at a 95% confidence level in the range of 0.8–1.8 myr, given that the mutations were random and fit a binomial distribution. No analysis can be made for the

TABLE 1
PCR primers, chromosomal locations, and PCR product sizes

Name	Accession	Position	5' primer sequence (5'-3')	3' primer sequence (5'-3')	A.T ^a	Human diversity ^a	Chr. ^c loc.	Product size ^c		
								Filled	Empty	411
Yb9NBC1	AC024091	26414-26105	ACTATCTTTAGATCCAGGGTGAAGC	TTCCAGTGGTAACTCTATGGAAAT	60	FP	12	411	86	
Yb9NBC2	AC024896	142649-142362	GCAGACAGTACCCACTTATTTTGT	TGGCTTCTATAAAGGATTTGTTCTTC	55	FP	7	462	146	
Yb9NBC3	AC005342	167963-167121	GAACCTCACTCTGCCATGTC	CATGGTCCGTCCTGCTTACA	60	FP	12	527	200	
Yb9NBC6	AC020900	61455-61742	CCAGACCCATATGTTCAATAAAATGAC	CCACCTGGAAAACACCCAAA	55	FP	5	493	153	
Yb9NBC7	AC009062	351148-351457	CACTAAATGATGCCAACACCTTC	CTAAATGTCAGCTATGCCACAGA	58	HF	16	403	83	
Yb9NBC8	AC011967	156726-157013	TAACTTTACTTTCATCGGCACATT	ACACTAGTTACCCCTCTCAGCAC	60	LF	18	419	86	
Yb9NBC9	AC022199	71329-71616	ACCTTCCCATTTTCGGTTGCTCTAT	CCCCTGTTAACCCAAACCTTT	60	FP	17	453	120	
Yb9NBC10	AC025961	22060-21773	GTITTCCTGGTGTGCCCTAAATA	TTAACCTTAACCTCACAAAGACCCAAAG	60	IF	4	524	197	
Yb9NBC11	AC019189	172700-172987	AAAGACTTTTCAGGTCTGTGAGCA	ATGCCATGTCCTATGCCAAACTATCAA	55	FP	1	392	74	
Yb9NBC12	AC011170	158821-159108	AACACCTGAGAAAGGCTCATTC	CCITGGAAAATACCTAACAAAGCAC	55	LF	10	414	93	
Yb9NBC13	AC003985	36492-36792	TCTAGTTGGAGTGGCCATGC	CITCCCACTCATGCTTCTCTGT	60	FP	7	510	167	
Yb9NBC15	AC006036	24671-25059	CCAAAGCTTTCAGCTTTATGCTC	CCTCAAAACCCGCTGAATTGT	60	FP	7	489	159	
Yb9NBC16	AC024057	13965-13678	GAACAAAGAATGCCAAAGGTAGTTG	ACCCCCATGACACTAAATTACCTAT	55	—	3	416	117	
Yb9NBC17	AC012664	22598-22811	ACTTACCCAAAACGCATGCATTC	AACGTAGATGCCAGACAAACACTCTT	60	FP	2	709	391	
Yb9NBC18	AC005751	35058-35343	CTTTTGAAGCTTACGTTTACCC	TCCCATGAGCTTACGTTGATCAT	60	FP	16	531	203	
Yb9NBC21	AL136081	33392-33083	TTCATGTAAGCCAAAACACTACCTGTT	TTAACAGCCTTACAGTTGGCAGAG	52	FP	6	425	107	
Yb9NBC22	AL139193	160587-160874	TGCCACAACCTATACACCAAGACACTG	TCTCTCCCATCATGAGAAACCTAACG	55	LF	14	435	110	
Yb9NBC23	AL356756	80321-80034	CAGGACTTTTATTGAAATCCTCACCT	AAACAGAGACATGCCCAATTAA	58	FP	14	412	83	
Yb9NBC25	AC005558	90385-90099	CAGITGTCAAAATTGCAATGGATAC	ACATCATTAAGCTCTCTCTGACATT	55	FP	5	496	159	
Yb9NBC27	AC011966	13523-13810	CATGGATAACATCAAAAGGCTTCAG	ACACTAGTTTACCCCTGTCACAC15	55	LF	15	482	149	
Yb9NBC28	AC004808	25856-26143	AAAAAGGTTGCCCTCCATATA	CTGTGCCCATAACTCAAACTGTAATG	55	FP	7	539	208	
Yb9NBC29	AC005008	29772-30089	GTAATATGAGCTGGATGGGGTTACT	GCTGAAAAGAAGAACCCCTAACTTAT	60	LF	7	474	138	
Yb9NBC30	AC003003	35922-36209	GAAGCCCCATCCATTCTCTTACA	CTGGCAAAATAATTGGCCGACT	60	IF	16	508	156	
Yb9NBC31	AF107258	58225-58541	TTTCTCTGGCAACTTCCCTGT	GACAGTGTACGGTACCTACC	56	FP	21	457	130	
Yb9NBC32	AL121582	154486-154199	CCTAACCCCTCATTTTACATTTC	GTCATTGCACTGTCAAAAGACTGT	55	FP	20	469	141	
Yb9NBC34	AL121841	28487-28800	CCAAACTTTCTCTGCTGCCAA	CACAAATACTCCCTGGCTCA	55	FP	14	489	90	
Yb9NBC35	AC040906	166712-167029	TTAACAGCTTACAGTTGGCAGAG	TCATCATGAGCCAAAACACTACCTCTTC	60	FP	6	427	109	
Yb9NBC36	AF015725	15626-15909	AAGGAGTCATCATCCATT	ACACACAAATACTCCCTACATT	60	FP	7	521	201	
Yb9NBC37	AB014460	3311-3621	CAAATGCGCCCTGTTCTCTT	GTCTCCACCGCATCTTGGCAG	62	FP	8/16	458	142	
Yb9NBC39	AC004542	12799-13104	AGAGGATCTTGGCAACT	GTCCTGTGGGGTTAGGAAGA	55	—	22	509	176	
Yb9NBC40	AP000237	60117-60404	ACTGAGTGGCACTACCCAAAT	CTTCAGGACTTACCCCTCTTACAT	60	FP	21	450	124	
Yb9NBC41	AC004140	4672-4851	TGTTTCCCTCATCTGCCAAACTT	AAAACACTGTCATGACCACTCAG	55	FP	7	761	389	
Yb9NBC42	AC004945	152516-152833	CACATCUCCTCCCTTCTCT	CAAAACCTGAAACATGGCTAA	55	FP	7	521	177	
Yb9NBC44	AC006561	13793-13506	CACTACAAACATACCATCCTCAA	GTATACGAAACAGCTGTGTCAC	55	FP	12	426	106	
Yb9NBC45	AL121978	17555-17288	GGAGAACCACTGAAACATGGAG	AGCCCTGGTATCCAGCTCTT	60	—	6	486	167	
Yb9NBC48	Z95114	30489-30202	GCTGCCATACCAACCCCTTGTGTC	TTCTGGCTGTAAGGCTCAGTAGG	60	FP	22	432	117	
Yb9NBC49	AC005375	129358-129604	ATCCCTTATAGATCACAGGTCAAC	CAACAACAAATCTGCTTCTGTCAC	58	FP	17	393	134	
Yb9NBC50	AL109865	28485-28772	GTTCCACAACTACACCGACAAATGT	GAAGCTCTTGTAGGAAACCAATCTC	55	IF	11	460	138	

* (continued)

TABLE 1
(Continued)

Name	Accession	Position	5' primer sequence (5'-3')		3' primer sequence (5'-3')		A:T ^a	Human diversity ^b	Chr. loc.	Product size ^c
			Filled	Empty	Filled	Empty				
Y _c 9NBC53	AQ382257	185-472	GGCACTCCCTATAATGACCTC	TAGGATGAAATGAACTTTCAGATG	CCACCAATTCTACCTGTATGG	HF	55	454	68	
Y _c 9NBC54	AL050305	39695-40005	CTCAGATAAGAAACTGAAACACAG	CCTATTTATAACCAATGACGAAACAG	CCATTATAACCAATGACGAAACAG	FP	58	492	172	
Y _c 9NBC55	AQ076355	91-379	TTGACTGTAACTGTTATTTTC	CCTATACCTTAAACAAAGCTTCGAC	CCTATACCTTAAACAAAGCTTCGAC	FP	60	425	108	
Y _c 9NBC58	AC022199	109995-109711	GTTCCTCTAGTCTCTGCATTTCG	TGACTAGTCCTTGTATTGAGAA	TGACTAGTCCTTGTATTGAGAA	FP	60	17	445	128
Y _c 9NBC59	AL121582	149776-150063	AGTACACTGAGTTCTATGCCCTC	CGCTGCCAGACACAAACACTT	CGCTGCCAGACACAAACACTT	FP	55	20	480	160
Y _c INBC1	AC011296	4067-3787	TCTCTCATGAAACATAGATAAAA	GATTTGTCCTAAATGCCCTAACT	GATTTGTCCTAAATGCCCTAACT	IF	60	?	481	159
Y _c INBC2	AC006195	139237-139517	GGATAACCCCTGGAAAAAGA	CGTGGCATCTTGTGAGATAAAAT	CGTGGCATCTTGTGAGATAAAAT	IF	60	7	443	102
Y _c INBC3	AC010072	48921-49201	CAAACCTGTGACCTTGACA	GAACACCACATGTAACCCCTCACC	GAACACCACATGTAACCCCTCACC	FP	63	14	405	92
Y _c INBC6	AC004016	82266-81986	CATCAAACCCCCACACACTCA	CACTGGCATTTAGGATTTGG	CACTGGCATTTAGGATTTGG	FP	65	7	1009	677
Y _c INBC8	AC007298	28402-28682	GGCCACCTCTGAAATAGCTT	TCCCTGGAGGCCACATGTTT	TCCCTGGAGGCCACATGTTT	FP	63	12	463	115
Y _c INBC9	AL121603	31558-31838	GGGAATGTTCTAGGACATGG	AGAAAATTCTGCTCATGTTCTCAG	AGAAAATTCTGCTCATGTTCTCAG	IF	63	14	490	159
Y _c INBC11	AF123462	93456-93176	TCCCCAACTCCCCATGCCCTTAA	TCCAACATGCCAGAAACAGA	TCCAACATGCCAGAAACAGA	FP	63	14	778	437
Y _c INBC13	AL122006	69774-70054	TTAGGGTCCCTGTGACTTGTGT	GGCCATTCCCTCACCAGCCATT	GGCCATTCCCTCACCAGCCATT	FP	60	1	504	165
Y _c INBC14	AL031734	146718-146998	CGGAATCGGCATACCAACTGCA	TGGCAAAAGCCATTTCCTCAAAG	TGGCAAAAGCCATTTCCTCAAAG	FP	60	1	464	149
Y _c INBC15	AL031650	85392-85112	GGAAATGTTCTAGGACATGG	ACCAAATGAAAGGGGAGACA	ACCAAATGAAAGGGGAGACA	FP	63	20	418	112
Y _c INBC20	AP001696	246018-246298	GGAAAGTAATGAAAGCATTCTAGG	AGAGGAGCCCTATTTCTT	AGAGGAGCCCTATTTCTT	FP	60	21	486	163
Y _c INBC23	AC004626	28992-29271	TCCTCAGGTAGGGATGTTAATGC	TTCTGAGCTGCTGGGGACT	TTCTGAGCTGCTGGGGACT	IF	60	16	445	120
Y _c INBC24	AL137013	69380-69041	TCAAAGGGGAATACTGGGGAAA	CGGGCAAAATGAACTAACAGTGGAA	CGGGCAAAATGAACTAACAGTGGAA	FP	60	1	408	88
Y _c INBC25	AC018637	72620-72340	GGGCAGGGGATCTAGGCCACT	TGCCCCCTGTTCATCTGTGC	TGCCCCCTGTTCATCTGTGC	FP	60	7	432	108
Y _c INBC26	AC027279	127822-128102	TCACITTCAGAAAGGGAAAAAA	TGTGCTGCTGCCCTGGACTCTCAA	TGTGCTGCTGCCCTGGACTCTCAA	FP	60	16	472	165
Y _c INBC28	AC017019	30139-29859	TGCTGAGCTTCTGGCTCTGGCTG	TGCTCTACCTCTTGGCTCACAC	TGCTCTACCTCTTGGCTCACAC	FP	60	—	Y	414
Y _c INBC30	AL157756	37868-38148	CCCCCTAAGCCCTTGTGTAAA	CAAAGTCATCTCTGACCCCCCAA	CAAAGTCATCTCTGACCCCCCAA	FP	60	14	497	177
Y _c INBC31	AC008062	103843-108563	TTCTGTAAAAGCCTTAACTCTCCA	CACCATTTCTCACTGTCACTTGTGC	CACCATTTCTCACTGTCACTTGTGC	IF	60	7	443	110
Y _c INBC32	AC005866	37960-37680	CGGAGGGCAACCGAGACAATAA	CGGGAGGTCTCACCCCTTCAA	CGGGAGGTCTCACCCCTTCAA	FP	60	12	425	114
Y _c INBC33	AL132994	40508-40788	CTTATGGCTTACACTAGAA	TCCATATCTAGCCTCTGATTC	TCCATATCTAGCCTCTGATTC	FP	60	14	500	186
Y _c INBC34	AL136382	87933-87653	CCCCAAACCCCTTCCAGAC	CAACCAACCTGGATGGAGTGG	CAACCAACCTGGATGGAGTGG	FP	60	—	1	477
Y _c INBC35	AC004638	32778-33058	CCGATTCTCCATGCCGTGAT	TCCAAGGCATTGGGGATACAA	TCCAAGGCATTGGGGATACAA	FP	60	16	481	162
Y _c INBC36	AL121903	24409-24129	CACAGGAACATTCCTCCACAA	CGGAAATTCCTGAAGGAAAAACTGG	CGGAAATTCCTGAAGGAAAAACTGG	FP	60	20	437	101
Y _c INBC37	AL049562	25982-25702	CGTGGCATTCCTCTCATCACA	GGCACTTTACCTAAAGACCTTCA	GGCACTTTACCTAAAGACCTTCA	FP	60	14	406	88
Y _c INBC38	AC000118	10509-10789	TCCAAAACCTCTCTTGGCAT	TGAAGGAAATTATGGGACTCTG	TGAAGGAAATTATGGGACTCTG	FP	60	7	435	113
Y _c INBC39	AP001695	126848-127128	TCCAGGAACCAAAACACAATTCA	TCAACCCCAACTCTGACCTCTCAA	TCAACCCCAACTCTGACCTCTCAA	FP	60	—	21	623
Y _c INBC45	AF218891	1964-1684	TGGCCACATTGGAAATTCAAACAT	TTCCTGCTGCTTAAGTCACACATGA	TTCCTGCTGCTTAAGTCACACATGA	FP	60	19	413	94
Y _c INBC46	ZB6061	56824-57103	CTTTGAAACCATGCAACCAAAGG	CAGTTTCCAACTTCTAGGGACTTCA	CAGTTTCCAACTTCTAGGGACTTCA	FP	60	5	406	101
Y _c INBC48	AC007094	66892-66612	TTCACCCACAAATTAACTGAAAGCTT	CAAACGGCACACAGCATCTGA	CAAACGGCACACAGCATCTGA	FP	60	—	7	521
Y _c INBC49	AC011493	52071-51791	TGTCATGCTTACTATGGACCCCTAC	CTCGGGGAGACATCCCCCTTCC	CTCGGGGAGACATCCCCCTTCC	FP	60	7	700	392
Y _c INBC50	AC010382	50258-49978	GGTATGGCCCAAAATTAACTCA	TCCAAGAGAACAGCAACCTACAGA	TCCAAGAGAACAGCAACCTACAGA	IF	60	5	406	101
Y _c INBC51	AC009415	123638-123918	TCAATACAAAAGACAGGCTTTC	CAAGGAAACAGATTCAAGAAAGACA	CAAGGAAACAGATTCAAGAAAGACA	FP	62	7	429	111
Y _c INBC52	AC002429	141029-140749	GCTTTGACACATCCCCACGT	CACAAGATTTGGGGCCAAGAC	CACAAGATTTGGGGCCAAGAC	FP	62	7	429	111
Y _c INBC53	AC004848	43020-42740	AAAGCTATCAACCATGCCAACAA	CAAATGCTTATTTGGGGAATCG	CAAATGCTTATTTGGGGAATCG	FP	62	7	505	186

(continued)

TABLE 1
(Continued)

Name	Accession	Position	5' primer sequence (5'-3')	3' primer sequence (5'-3')	A.T ^a	Human diversity ^b	Chr. ^c loc.	Product size ^d
Yc1NBC56	AC006017	155231-154951	TCTGTAAAAACTGCCCTCACAT	GGGGCTGTGATATTCTGTGCTG	55	—	7	593 287
Yc1NBC58	AL133367	83515-83795	TGCTGCCATCAATCAGCCAGA	TCCCACTTCTGGCAACCAT	65	FP	14	427 118
Yc1NBC59	AC006213	58483-58763	ACCCCTCCCTCCCTCTGTGG	CCCTGGCAAAACGCTGGAAA	60	FP	19	428 93
Yc1NBC60	AL136319	30378-30658	CAAACCCCCAAGATTCTCACCC	TCTCCATCATGATTCCCAACTGA	60	IF	10	522 205
Yc1NBC63	AL121964	57668-57943	GCTACTCACTAACACATCAAGA	AAGCTGGCTGGCTGGCTTCAC	60	IF	6	502 181
Yc1NBC64	AL121904	25022-25262	CACATCTGGTCTAGGAGCTG	CAAGCTGTGTATTCTTGATCTGC	60	IF	20	600 292
Yc1NBC65	AL049643	46216-45936	TTGGCTGACCATATCAGATCTG	TCCAGTGCTTAAGACATAAGCAAGC	60	FP	X	456 152
Yc1NBC66	AJ006998	11416-11136	GGCTTAGCAACGCTCTTITGC	TGATGACTGTACAAGGCCACACTT	60	FP	21	422 110
Yc1NBC69	AB020859	19030-19310	CCACATTTATCAGTACCTACA	CCITGCCAGAATAGCAATGTAT	55	IF	8	524 210
Yc1NBC70	AL133238	24939-24661	ACCAATTGTCGAGGCCACAA	GACGGTGCCTTAGTGGGAGCAA	60	FP	14	452 137
Yc1RG60	AC019215	161766-162046	TCCCACATTTCAGTGTGAATT	GGCATTTGGATAGTTCCTCTG	60	HF	8	474 159
Yc1RG62	AC007428	139021-139301	GCTCAACATGCCATAACCTGAAC	ATITCCAAAGAAACCCCTGACT	60	FP	?	522 216
Yc1RG83	AC009004	751-1030	CTGGCTGCCAGATTGTTAAA	GTGGAACACGTGATTGCTGTGA	60	FP	19	724 397
Yc1RG64	AC009289	65992-66272	TCCACTCATCTTAATGTCGCTTAG	CCATACACCTTCCCTTCTGATT	60	FP	14	380 67
Yc1RG65	AC019181	63269-63549	GCAGCCCTGGCATATCAATTAGG	ATCGGTAAAACCTCTAGCACTG	60	FP	2	735 413
Yc1RG66	AC009506	178981-179192	CTTTTCTCAGACTGTGCTTGC	GCAACAAGACAAAAGCAACCTG	60	FP	1	419 109
Yc1RG67	AC008039	164672-164954	AAAATACCTTCCCAGACTCC	CCCTAACGGACTTTATAATGGCACT	60	FP	7	382 125
Yc1RG68	AC008039	164672-164954	ATGGCTGTCCACAGAAACTGAG	GCAGGGCTCCATTATACTGCTCTG	60	IF	7	480 166
Yc1RG70	AC006323	3461-3741	CTCTGCCACCATGACAAATCAAT	CACCATCTAAAGCACTCACTICA	60	FP	17	504 178
Yc1RG71	AC011450	98261-98574	TACTGAAGACCAACTGGCCACAA	TTCCACTCACCTTACCCAGATA	60	FP	19	435 73
Yc1RG73	AC007739	154145-154426	ATTGGCCAAGAACCTTGTGTTTC	CCGGTTGAGAAAAGTCAGTG	60	FP	?	463 143
Yc1RG74	AC006038	73850-74014	AACTACCCCTGAAATGGGCAAATAAT	GAATGGATTCGAAACCAACATAA	60	FP	2	415 226
Yc1RG77	AC005783	19041-19327	GAGAAAGAGCTCCAAAGGATGTC	CAACTAACCCCCAAATGAGGT	60	FP	19	401 84
Yc1RG78	AC002044	13430-13712	CTCCAGGATCTGGTTTCATCTTA	TCATGGTAACTAGGACAAAGATCC	60	FP	16	431 119
Yc1RG79	AC004690	35856-36140	GGCTCTATCATCACCTTAATTTCGA	TGGTTRTTAGATGCCAAACACTAT	60	FP	7	497 158
Yc1RG80	AC004485	74445-74724	QAGAAATTGGCTCTTACAGTTTCC	AGACGTGAACAGCTTATTGCTCTGA	60	FP	7	482 134
Yc1RG81	AF088219	176700-176982	CACACAGCAGGATACAAAAAC	CTTCTAGGCTTTAGTTCGGGAAC	60	FP	17	535 354
Yc1RG82	AF088219	99726-100005	CCTGGACCTTCTAGGCTATT	CACTCATCTCATCTCACAGCAC	60	FP	17	388 91
Yc1RG83	AC005026	82038-82232	GGCACTATGCTGGCTGTATAG	CGAAAACCTTAATGCTTCCCTCT	60	FP	7	389 153
Yc1RG84	AF131217	50031-50317	CCACTCTCCACCTCTATGGTAT	AAAATGCCACAGGAATAGCGTTTC	60	FP	21	387 60
Yc1RG86	AC005412	78372-78652	ATGGGCTGACCACTTGTATTGAC	CTTCTGGAGGGGAACCTGTTTTA	60	FP	17	499 188
Yc1RG87	AC008071	84205-84487	GAACATCTGAAACACAGTGTAGG	CACTCATCTCAAGTCACACAGC	60	FP	7	427 92
Yc1RG88	AC006305	13802-14086	GGTGAACCTCAACCCCTAACCTCA	GTGCATTCCCGACACAAGTATT	60	FP	18	395 74
Yc1RG90	AC004671	68017-68298	CCCTTAATTAATTTCCTCCGGATT	GCTGTAGCTGGCTAAATACCAAC	60	FP	12	398 100
Yc1RG91	AC005288	37818-38107	AATGGGTGAAAAGAGCTAACACC	TGTGTAGCTTAACAAAGGATGG	60	FP	17	700 391
Yc1RG92	AC004675	78485-78767	ACACTCTATGGAGGCCAGTCATCT	CCTGGACCTTCTAGGCTTT	60	FP	17	402 85
Yc1RG93	AC005324	137294-137574	GGCATTCAACTGTCGCTAGAAT	AGGAAGGCCAATATGATGTCG	60	IF	17	377 63
Yc1RG95	AL049537	38717-38997	ACCTAACAGATGACTGCTGAAA	GACCTAGAGAAAGCAAGGATTC	60	IF	20	701 390
Yc1RG96	AF042091	61095-61379	ACACACAAACCTGAAAACCTCAACC	CCACACCACCATCTTATTGAT	60	FP	21	457 128

* (continued)

TABLE 1
(Continued)

Name	Accession	Position	5' primer sequence (5'-3')	3' primer sequence (5'-3')	A.T. ^a	Human diversity ^b	Chr. ^c	Product size ^d
						loc.	Filled	Empty
Yc1RG97	AF042090	42069-42352	AAGTCACACTTTCACGTTCAC	CCTTGATTGCCATTCACGTTTA	60	HF	21	441
Yc1RG98	U92032	3903-4188	TCCTATCTCTGACACCTTGACACG	AAAGAACCCAGACCTATGACAGA	60	FP	6	442
Yc1RG99	AL022163	85635-86116	AAAGCCTCTGGTACAGAACATCAGC	CCATGGCCGAAGTTAATGAGAAGT	60	IF	X	113
Yc1RG100	AL354872	86112-86401	ACITCCATGAGCTACTGGCTGTA	GATCTCTAAACCATTAAGGCTCAC	60	LF	1	64
Yc1RG101	AL031662	26328-26613	CCAGCCAAGACATTACCAAAA	CTCCAGTCATTCATTCAAAGAAG	60	HF	20	143
Yc1RG102	AL158040	201136-201376	CTGCCCTTGTAGTAATGCTCAAGG	GTAGACATTGGCTCCACCTTAT	60	FP	10	235
Yc1RG103	AL158157	101226-101505	GGCATTTGGCATTCTGATGCCITA	GACATGTTAGAGAAAAGGTACATC	60	HF	9	110
Yc1RG104	AL157384	87495-87786	CTGGAAGGGATCTTCTTATGCC	CCCTTTCTGATCTTATTCTCCA	60	FP	9	79
Yc1RG107	AL358293	139195-139492	GTITGATCAGCTGCTCTCAACT	TGAATGAAATTTCAGTTGCTGCA	60	FP	9	130
Yc1RG108	AL035458	29348-29629	CTTATATCAACAAACCCCCGTA	CACCAAAAGAACCGACAAAGAAC	60	FP	14	76
Yc1RG109	AL137794	36815-37094	GCTAGAAATTCAATAATGCCATCC	TCCAGTTGAGTTGCACTGATT	60	FP	20	71
Yc1RG110	AL109824	732-1012	CTAGGCTTAAGGACTCCCTGG	GTGACCTAGGCCAGAGGTTAATG	60	FP	1	188
Yc1RG113	AL163278	90774-91055	CTGTCACCGCTAAAGACCTTCTGTG	CATATCTCAGGCCAAATGGCAAC	60	FP	20	85
Yc1RG114	Z98051	36444-36724	ATCAGCCATACAGTCTGAAAGC	AATCTTGGTTAGTGTGACTCAACC	60	FP	21	76
Yc1RG115	Z98046	60991-61271	GTTCCTGCTCTTTGCTATCTGCAAT	GTGGTCAAGGTACAGACTCATCC	60	FP	X	110
Yc1RG116	AL078621	142330-142621	GGTTAAAAGAACACATGGATGG	GAAAGGTGGGTGCTAAATGCTA	60	FP	X	392
Yc1RG117	AL096861	42260-42540	CAATAACCCAAACCTGGTAGCTG	TGCAATAAACAGTGTCTCTCC	60	FP	22	419
Yc1RG118	Z71183	21436-21716	TACACAGACCAATGGCCAAACTA	TCCAGATCCATGACATAACACT	60	FP	X	166
Yc1RG120	AL023283	61027-61306	TCTGGCTCTTGGCTATAACACTGCTG	GAAGGGCACTGAATGACACACT	60	FP	22	389
Yc1RG121	AL109760	24171-24451	CATGGACATTTGGGAGAATGTA	GGCCCTATAATTACTCAGCAG	60	FP	6	194
Yc1RG122	AL023882	16690-16970	CACACACACACAAATTAGCC	GTCAGCTCTGAAACGGCTTAC	60	FP	4	398
Yc1RG124	AL022897	18401-18681	AAATCACTGTACCAACCCCTGTCA	GCAAAACACCACTGAACCATAA	60	LF	16	92
Yc1RG125	X76070	298-578	TGTTCCTCTGGCTCCATTC	CTGTTCTCTATGATCTGAAAGATGG	60	FP	1	234
Yc1RG126	AP001752	250076-250356	CCCTGTAGTAATGCCCTCACTGAA	GGCGATTAGGGATACACATAGA	60	IF	2	125
Yc2NRC1	AC002430	108794-109074	ACATAGTGGCCATCAAGAC	CTTAATGTTCCATTTCTCCA	55	FP	21	91
Yc2NRC5	AC007384	128277-128557	GAAGGAATACAGCCAGGAAT	CTCCCAAAACAACCTAAACCC	55	IF	7	131
Yc2NRC9	Z98051	36444-36724	AAAAAGCCTGATACCTTTCG	CTTGGTTAGTCTGACTCAACC	55	IF	7	125
Yc2NRC11	Z69666	9696-9416	CGACAAGTGAACCTTACCG	CTCCTCCCAATGATCTATGCT	55	FP	X	407
Yc2NRC13	AC007882	150095-149815	TGGGATAATGATTGTCTCC	AACATGTTGGCAGATGATCA	60	FP	16	91
Yc2NRC15	AC007541	129217-129497	GGTAAGGAAAACCAACTAA	CITTTGAGGAAGCTGATGAC	55	FP	12	82
Yc2NRC17	AC005541	74318-74593	ATCAAATGGCAGCCCTTACT	GCCTTTCATTCCTGACTTA	60	FP	7	92
Yc2NRC19	AL022163	81833-82113	GCTTAAAGCACTGGTACAGA	TGGCGAAGTTAATGAGAAGT	55	HF	X	67

Perfect matches to the consensus are in italics.

^aAmplification of each locus required 2 hr 30 min at 94° initial denaturing and 32 cycles for 1 min 94°, 1 annealing temperature (A.T.), and 1 min elongation at 72°. A final extension time of 10 min at 72° was also used.

^bAllele frequency was classified as fixed present (FP), low (LF), intermediate (IF), or high frequency (HF) insertion polymorphism. Fixed present: every individual tested had the Alu element in both chromosomes. Low frequency insertion polymorphism: the absence of the element from all individuals tested, except for one or two homozygous or heterozygous individuals. Intermediate frequency insertion polymorphism: the Alu element is variable as to its presence or absence in at least one population. High frequency insertion polymorphism: the element is present in all individuals in the populations tested, except for one or two heterozygous or absent individuals. —, indeterminable.

^cChromosomal location determined from accession information or by PCR analysis of NIGMS monochromosomal hybrid cell line DNA samples.

^dEmpty product sizes calculated by removing the Alu element and one direct repeat from the filled sites that were identified.

Y	GGCCGGGCGCGGTGGCTCACGCCGTAAATCCCAGCACTTTGGGAGGCCGAGGCAGGGCGGA	60
Yb8	T... 60
Yb9	T... 60
Y	TCACGAGCTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAACCCCGTCTACTAA	120
Yb8	...T.....	A 120
Yb9	...T.....	A 120
Y	AAATACAAAAAAATTAGCCGGCGGTGGTGGCGGGCGCTGTAGTCCCAGCTACTCGGGAGG	180
Yb8	C..... 180
Yb9	C..... G 180
Y	CTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGGAGCTTGCACTGAGCCGAGATCGCGC	240
Yb8	A T... 240
Yb9	A T... 240
Y	CACTGCCACTCCA-----GCCTGGCGACAGAGCGAGACTCCGTCTC	281
Yb8G...GCAGTCCG.....	288
Yb9G...GCAGTCCG.....	288

Yc1 and Yc2 Alu elements, because only subfamily members with perfect identity to the subfamily consensus sequence or one mismatch were isolated from the database using one of the database screening procedures.

Phylogenetic distribution and human genomic diversity of the new subfamilies: Amplification of the Yb9, Yc1, and Yc2 elements from nonhuman primate genomes facilitated the analysis of the phylogenetic distribution of these elements, using PCR and the oligonucleotide primers in Table 1. The majority of the elements evaluated were absent from the genomes of the nonhuman primates, suggesting that these elements dispersed and were fixed in the human genome after the human and African ape divergence.

We performed a PCR analysis on a panel of human DNA samples to determine the levels of human diversity associated with the Alu elements from these new subfamilies, using the oligonucleotide primers shown in Table 1. The panel consists of 20 individuals of European origin, African-Americans, Asians, and Egyptians for a total of 80 individuals (160 chromosomes). We were able to analyze 28 out of the 56 Yb9 elements, 97 out of 176 Yc1 elements, and 8 out of 17 Yc2 Alu elements, using this approach. Several factors did not allow for analysis of all the elements. Mainly, we were unable to design appropriate primers due to insufficient flanking unique DNA sequences or because the element analyzed resided within another type of repeat as described previously (BATZER *et al.* 1991). The Alu elements were classified as fixed present and high, intermediate, or low frequency insertion polymorphisms (see Table 1 for definitions). In general, we observed that approximately one-fourth to one-third of the elements analyzed had some degree of insertion polymorphism (Yb9 with 10/

FIGURE 1.—Consensus sequence alignment of Y, Yb8, and the potential new subfamily Yb9 identified. Nucleotide substitutions at each position are indicated with the appropriate nucleotide. Deletions are marked by dashes (-). The Yb8 and Yb9 diagnostic nucleotides are indicated in boldface type with the corresponding diagnostic numbers above.

28, Yc1 with 24/97, and Yc2 with 3/8). The population-specific genotypes and levels of heterozygosity for each element are shown in Table 2. The high proportion of polymorphic elements in these Alu subfamilies is in good agreement with our previous observations, indicating that these subfamilies are very recent in origin and still actively retroposing within the human genome.

DISCUSSION

From our subset of AluYb8 and Y elements, we were able to retrieve three Alu subfamilies termed Yb9, Yc1, and Yc2. A schematic of the evolutionary relationship of these subfamilies with the previously defined Alu subfamilies is shown in Figure 3. Alu subfamilies arise as a result of mutations occurring in an existing master element or new source elements capable of significant amplification. In this case, the new subfamilies are presumably examples of Alu subfamilies that may have originated from the rare instances when an Alu element fortuitously becomes both transcriptionally and retropositionally active, therefore allowing it to be another Alu source gene.

The young Alu subfamilies are currently active with respect to retroposition, whereas the older Alu subfamilies typically are not. The old Alu subfamilies (Sx, J, and Sg1), which comprise the vast majority (>1,000,000 copies) of the Alu elements present in the human genome, appear completely inactive as none of their members have been associated with *de novo* Alu inserts that result in human diseases (Table 3). When noting the ratio of reported Alu insertions associated with diseases and the estimated size of the Alu subfamily, the younger

Y	GGC GGGGCGCGGTGGCTACGCCTGTAATCCCAGC	ACTTTGGGAGGCCGAGGCGGGCGGA	60
Yc1	60
Yc2	60
Clinh	60
BRCA2	60
GK	T.....	60

Y	TCACGAGGT CAGGAGATCGAGACC	ATCCTGGCTAACACGGTGAAACCCCGTCTACTAA	120
Yc1	120
Yc2	A	120
Clinh	T.....	120
BRCA2	T.....	120
GK	120

Y	AAA -----TACAAAAAAATTAGCCGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCG	175	
Yc1	A	175
Yc2	A	175
Clinh	A	180
BRCA2	A	175
GK	A	175

Y	GGAGG CTGAGGCAGGAGAATGGCGTGAACCCGGAGGCCGAGCTTGCAGTGAGCCGAGAT	235	
Yc1	235
Yc2	235
Clinh	G	240
BRCA2	235
GK	G	235

Y	CGCGCC ACTGCACTCCAGCCTGGCGACAGAGCCAGACTCCGTCTC	281	
Yc1	281
Yc2	281
Clinh	286
BRCA2	281
GK	281

subfamilies Ya5, Yb8, and Yc1 currently appear to be ~1000 times more active than the Alu Y subfamily with 7/2640, 3/1852, and 3/400 compared to 1/200,000 (Table 3). The Alu Ya5a2 subfamily appears to have even a higher current retroposition rate (1/40), but the very young age and small size of the subfamily may be an influencing factor. In general, two independent observations support the current mobility of these young Alu subfamilies within the human genome. First, there are examples of Alu inserts that have caused disease that belong to these young subfamilies. Second, the subfamilies have a high proportion of Alu insertion polymorphisms between individuals/populations (Table 3), indicating the recent proliferative/amplification activity of these Alu elements in the human genome.

Alu elements that are polymorphic for insertion presence/absence have previously proven useful for the study of human population genetics and forensics (BATZER *et al.* 1991, 1994; PERNIA *et al.* 1992; NOVICK *et al.* 1993; HAMMER 1994; TISHKOFF *et al.* 1996; STONEKING *et al.* 1997; MAJUMDER *et al.* 1999; COMAS *et al.* 2000; JORDE *et al.* 2000; WATKINS *et al.* 2001). The identification of

very young Alu subfamilies with a high proportion of polymorphic members provides new sources of Alu insertion polymorphisms for the study of human population genetics. However, it is important to note that an exhaustive analysis of these small subfamilies will only generate a relatively small number of new Alu insertion polymorphisms.

Master element vs. source gene: Alu elements have been proposed to fit an evolutionary model where the copies arose from "master" genes (DEININGER and SLAGEL 1988; LABUDA and STRIKER 1989; SHEN *et al.* 1991; DEININGER *et al.* 1992). A master gene can be defined as an element that is highly active during a long period, therefore generating a lot of copies of itself. However, we demonstrated that recently inserted Alu elements (*de novo*) belong to a variety of Alu subfamilies, indicating the simultaneous presence of multiple active elements in the human genome. These active elements that have a low rate of amplification and are only active for a very short period of time should not be classified as master genes. To distinguish between them, we suggest the use of the nomenclature of "master gene" when

FIGURE 2.—Consensus sequence alignment of Y, Yc1, Yc2, and three Alu Yc1 elements associated with disease. The diseases linked with Yc1 Alu elements are the angioedema caused by a *de novo* insertion in the C1 inhibitor gene (Clinh; STOPPA-LYONNET *et al.* 1990), breast cancer with another *de novo* insertion in BRCA2 (BRCA2; MIKI *et al.* 1996), and glycerol kinase deficiency (GK; ZHANG *et al.* 2000). Nucleotide substitutions at each position are indicated with the appropriate nucleotide. Deletions are marked by dashes (-). The diagnostic nucleotides are indicated in boldface type with the corresponding diagnostic numbers above.

Alu Insertion Polymorphism

TABLE 2
Alu Yb9, Yc1, and Yc2 associated human genomic diversity

Elements	African American			Asian/Alaska native			European			Egyptian		
	Genotypes			Genotypes			Genotypes			Genotypes		
	+/+	+-	-/-	fAlu	Het ^a	-/-	fAlu	Het ^a	-/-	fAlu	Het ^a	-/-
Yb9NBC8	0	0	20	0.000	0.000	0	19	0.000	0.000	0	17	0.000
Yb9NBC7	19	0	0	1.000	0.000	19	0	1.000	0.000	0	3	0.925
Yb9NBC10	3	1	4	0.438	0.525	2	0	14	0.125	0.226	3	0
Yb9NBC12	1	6	12	0.211	0.341	0	14	5	0.368	0.478	0	14
Yb9NBC22	0	0	14	0.000	0.000	0	0	15	0.000	0.000	9	8
Yb9NBC27	0	0	15	0.000	0.000	0	0	12	0.000	0.000	0	15
Yb9NBC29	0	1	9	0.050	0.100	0	7	12	0.184	0.309	0	2
Yb9NBC30	2	1	11	0.179	0.304	0	3	16	0.079	0.149	0	6
Yb9NBC50	0	0	15	0.000	0.000	0	6	7	0.231	0.369	1	0
Yb9NBC53	13	0	2	0.867	0.239	20	0	0	1.000	0.000	15	0
Yc1NBC1	1	7	12	0.225	0.073	0	2	18	0.050	0.062	0	10
Yc1NBC2	1	13	6	0.375	0.038	0	15	5	0.375	0.038	1	15
Yc1NBC9	4	13	3	0.525	0.008	3	13	4	0.475	0.008	3	8
Yc1NBC23	0	0	18	0.000	0.000	0	0	19	0.000	0.000	0	18
Yc1NBC31	0	0	18	0.000	0.000	0	0	19	0.000	0.000	0	0
Yc1NBC35	1	6	7	0.286	0.073	2	10	8	0.350	0.045	0	0
Yc1NBC50	0	2	18	0.050	0.062	14	4	0	0.889	0.081	4	9
Yc1NBC51	0	4	18	0.091	0.169	0	0	18	0.000	0.000	0	20
Yc1NBC53	8	7	1	0.719	0.070	3	12	1	0.563	0.022	1	13
Yc1NBC60	6	9	3	0.583	0.027	6	9	5	0.525	0.008	5	11
Yc1NBC63	0	0	—	—	—	1	5	8	0.250	0.082	3	6
Yc1NBC64	0	0	5	0.000	0.000	0	5	8	0.192	0.323	0	5
Yc1NBC69	0	0	13	0.000	0.000	8	4	5	0.588	0.030	2	4
Yc1RG60	16	0	4	0.800	0.328	19	0	0	1.000	0.000	14	5
Yc1RG68	1	4	14	0.158	0.273	6	6	8	0.450	0.508	3	7
Yc1RG93	0	0	20	0.000	0.000	0	0	20	0.000	0.000	0	19
Yc1RG95	2	17	1	0.525	0.512	4	15	0	0.605	0.491	0	20
Yc1RG97	19	1	0	0.975	0.050	19	0	0	1.000	0.000	20	0
Yc1RG99	19	1	0	0.975	0.050	6	14	0	0.650	0.467	8	11
Yc1RG100	0	0	18	0.000	0.000	0	0	19	0.000	0.000	0	18
												0
												16

(continued)

TABLE 2
(Continued)

Elements	African American				Asian/Alaska native				European				Egyptian						
	Genotypes	+/-	+/+	-/-	fAlu	Genotypes	+/-	-/-	fAlu	Genotypes	+/-	-/-	fAlu	Genotypes	+/-	-/-	fAlu	Genotypes	
<i>Yc1RG101</i>	20	0	0	1,000	0,000	17	0	0	1,000	0,000	17	2	0	0,947	0,102	16	0	1,000	0,000
<i>Yc1RG103</i>	16	2	2	0,850	0,262	19	0	0	1,000	0,000	18	0	0	1,000	0,000	15	0	1,000	0,000
<i>Yc1RG123</i>	0	0	20	0,000	0,000	0	0	0	20	0,000	0,000	0	0	20	0,000	0,000	0	20	0,000
<i>Yc1RG125</i>	0	16	4	0,400	0,492	0	9	11	0,225	0,358	0	17	3	0,425	0,501	0	19	0	0,500
<i>Yc2NBC1</i>	1	4	3	0,375	0,061	3	6	5	0,429	0,027	13	1	0	0,964	0,065	10	3	1	0,821
<i>Yc2NBC5</i>	3	10	4	0,471	0,010	3	10	1	0,400	0,031	17	3	0	0,925	0,071	13	4	0	0,882
<i>Yc2NBC19</i>	15	3	0	0,917	0,077	18	0	0	1,000	0,000	7	10	3	0,600	0,03	14	4	1	0,842

^aThis is the unbiased heterozygosity.

^bAverage heterozygosity is the average of the population heterozygosity.
Elements in italics were screened using DNA collected from Alaska natives rather than from the Asian population.

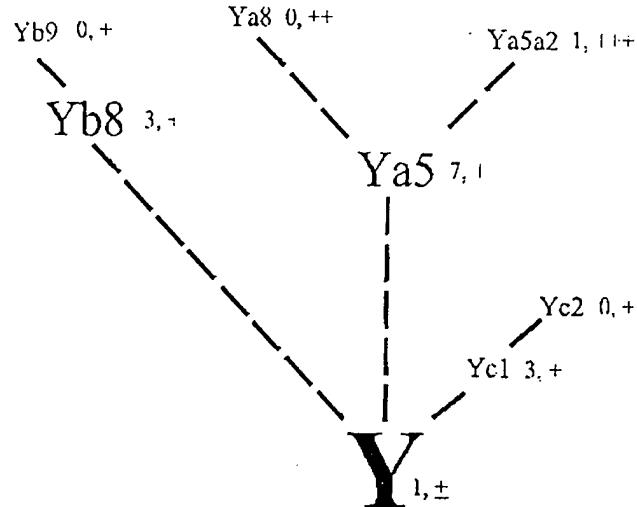


FIGURE 3.—Schematic diagram of the evolution of recently integrated Alu subfamilies. All the origins of the young Alu subfamilies are shown. The origins of the *Yb9*, *Yc1*, and *Yc2* Alu subfamilies are shown after the divergence of the *Yb8* and the *Y* subfamily, respectively. The size of the font is relative to the number of elements within each subfamily, the largest representing 100,000–200,000 copies; medium, 1000–2000 copies; and the smallest, 50–500 copies. The total number of elements from each subfamily linked to disease is indicated to the right. The proportion of polymorphic elements within each family is represented by the following: ±, rarely polymorphic elements are found; +, low percentage of polymorphic elements; ++, ~50% the elements are polymorphic; and ++++, most of the elements are polymorphic.

referring to the highly active genes for long evolutionary periods of time, like the Alu element that generated the majority (>90%) of the Alu elements currently present in the genome today. For those copies, or daughters, that acquired the ability to retropose we propose the use of the term “source genes.” However, some of the elements classified as source genes may be potential master genes, and only the progression of time will allow the appropriate distinction to be made.

Evolutionary reduction in the Alu retroposition rate: Our data indicate the existence of several currently active Alu elements that belong to different subfamilies within the human genome. However, the present amplification rate of Alu elements has drastically decreased from when it reached its peak 35 and 60 million years ago (mostly *Sx* subfamily). The majority of the Alu elements present in the genome of extant humans inserted during this peak amplification period. There are multiple reasons that could explain the reduction in the amplification rate of Alu elements. First, mutations within or near the master Alu element could reduce its retroposition activity or even totally abolish it by a variety of mechanisms (DEININGER and BATZER 1993; SCHMID 1996). Alternatively, mutations within the master gene or in the LINE elements that affect the ability to “parasitize” LINE element-encoded enzymes necessary for retroposition could also reduce the Alu amplification rate. Furthermore, the host may have also evolved cellular

TABLE 3
Young Alu subfamilies copy number, inserts linked to disease, and polymorphism

Alu subfamily	Estimated copy number	Inserted linked with disease ^a	General subfamily polymorphism ^b (%)
J, Sx, Sg1	>1,000,000	0	—
Y	>200,000	1	±
Ya5	2640	7	+
Ya5a2	40	1	+++ 80 ^c
Ya8	70	0	++ 50
Yb8	1852	3	+
Yb9	80	0	+
Yc1	400	3	+
Yc2	ND	0	+
			37.5 ^c

ND, not determined.

^a Previously published Alu elements linked with disease (DEININGER and BATZER 1999).

^b The proportion of polymorphic elements within each family is represented by the following: ±, rarely polymorphic elements are found; +, low percentage of polymorphic elements; ++, ~50% the elements are polymorphic; and +++, most of the elements are polymorphic.

^c Percentage polymorphism was determined using a selected subgroup introducing a bias.

mechanisms to reduce Alu proliferation. Finally, the availability of suitable genomic "insertion sites" may be reduced, since most evolutionarily neutral or positive sites are presumably already "filled" with different types of preexisting repeats. Alternatively, new Alu insertions may result in unacceptable local levels of unequal homologous recombination (DEININGER and BATZER 1999).

AMR was supported by a Brown Foundation fellowship from the Tulane Cancer Center. This research was supported by National Institutes of Health RO1 GM45668 (P.L.D.); Department of the Army DAMD17-98-1-8119 to (P.L.D. and M.A.B.); Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05)-05 and HEF (2000-05)-01 (M.A.B. and P.L.D.); and award 1999-IJ-CX-K009 from the Office of Justice Programs, National Institute of Justice, Department of Justice (M.A.B.). Points of view in this document are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice.

LITERATURE CITED

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.

AUSUBEL, F. M., R. BRENT, R. E. KINGSTON, D. D. MOORE, J. G. SEIDMAN *et al.*, 1996 *Current Protocols In Molecular Biology*. John Wiley & Sons, Canada.

BATZER, M. A., and P. L. DEININGER, 1991 A human-specific subfamily of Alu sequences. *Genomics* **9**: 481-487.

BATZER, M. A., V. A. GUDI, J. C. MENA, D. W. FOLTZ, R. J. HERRERA *et al.*, 1991 Amplification dynamics of human-specific (HS) Alu family members. *Nucleic Acids Res.* **19**: 3619-3628.

BATZER, M. A., M. STONEKING, M. ALEGRIA-HARTMAN, H. BAZAN, D. H. KASS *et al.*, 1994 African origin of human-specific polymorphic Alu insertions. *Proc. Natl. Acad. Sci. USA* **91**: 12288-12292.

BATZER, M. A., C. M. RUBIN, U. HELLMANN-BLUMBERG, M. ALEGRIA-HARTMAN, E. P. LEEFLANG *et al.*, 1995 Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *J. Mol. Biol.* **247**: 418-427.

BATZER, M. A., P. L. DEININGER, U. HELLMANN-BLUMBERG, J. JURKA, D. LABUDA *et al.*, 1996 Standardized nomenclature for Alu repeats. *J. Mol. Evol.* **42**: 3-6.

COMAS, D., F. CALAFELL, N. BENCHESI, A. HELAL, G. LEFRANC *et al.*, 2000 Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. *Hum. Genet.* **107**: 312-319.

DEININGER, P. L., and M. A. BATZER, 1993 Evolution of retroelements, pp. 157-196 in *Evolutionary Biology*, edited by M. K. HECKHT, *et al.* Plenum Publishing, New York.

DEININGER, P. L., and M. A. BATZER, 1999 Alu repeats and human disease. *Mol. Genet. Metab.* **67**: 183-193.

DEININGER, P. L., and V. SLAGEL, 1988 Recently amplified Alu family members share a common parental Alu sequence. *Mol. Cell. Biol.* **8**: 4566-4569.

DEININGER, P. L., M. A. BATZER, C. A. HUTCHISON and M. H. EDGEWELL, 1992 Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**: 307-311.

HAMMER, M. F., 1994 A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. *Mol. Biol. Evol.* **11**: 749-761.

JORDE, L. B., W. S. WATKINS, M. J. BAMSAD, M. E. DIXON, C. E. RICKER *et al.*, 2000 The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am. J. Hum. Genet.* **66**: 979-988.

JURKA, J., 1995 Origin and evolution of Alu repetitive elements, pp. 25-42 in *The Impact of Short Interspersed Elements (SINEs) on the Host Genome*, edited by R. J. MARAIA. R. G. Landes Company, Austin, Texas.

LABUDA, D., and C. STRIKER, 1989 Sequence conservation in Alu evolution. *Nucleic Acids Res.* **17**: 2477-2491.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

MAJUMDER, P. P., B. ROY, S. BANERJEE, M. CHAKRABORTY, B. DEY *et al.*, 1999 Human-specific insertion/deletion polymorphisms in Indian populations and their possible evolutionary implications. *Eur. J. Hum. Genet.* **7**: 435-446.

MIKI, Y., T. KATAGIRI, F. KASUMI, T. YOSHIMOTO and Y. NAKAMURA, 1996 Mutation analysis in the BRCA2 gene in primary breast cancers. *Nat. Genet.* **13**: 245-247.

MIYAMOTO, M. M., J. L. SLICHTOM and M. GOODMAN, 1987 Phylogenetic relations of humans and African apes from DNA sequences in the psi η -globin region. *Science* **238**: 369-373.

NOVICK, G. E., T. GONZALEZ, J. GARRISON, C. C. NOVICK, M. A. BATZER *et al.*, 1993 The use of polymorphic Alu insertions in human DNA fingerprinting. *Exper. Suppl.* **67**: 283-291.

PERNA, N. T., M. A. BATZER, P. L. DEININGER and M. STONEKING, 1992 Alu insertion polymorphism: a new type of marker for human population studies. *Hum. Biol.* **64**: 641-648.

ROY, A. M., M. L. CARROLL, D. H. KASS, S. V. NGUYEN, A.-H. SALEM *et al.*, 1999 Recently integrated human Alu repeats: finding needles in the haystack. *Genetica* **107**: 149-161.

ROY, A. M., M. L. CARROLL, S. V. NGUYEN, A.-H. SALEM, M. OLDRIDGE *et al.*, 2000 Potential gene conversion and source gene(s) for recently integrated Alu elements. *Genome Res.* **10**: 1485-1495.

SCHMID, C. W., 1996 Alu: structure, origin, evolution, significance and function of one-tenth of human DNA. *Prog. Nucleic Acid Res. Mol. Biol.* **53**: 283-319.

SHEN, M. R., M. A. BATZER and P. L. DEININGER, 1991 Evolution of the master Alu gene(s). *J. Mol. Evol.* **33**: 311-320.

SMIT, A. F., 1996 The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6**: 743-748.

STONEKING, M., J. J. FONTIUS, S. L. CLIFFORD, H. SOODYALL, S. S. ARCOT *et al.*, 1997 Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* **7**: 1061-1071.

STOPPA-LYONNET, D., P. E. CARTER, T. MEO and M. TOSI, 1990 Clusters of intragenic Alu repeats predispose the human C1 inhibitor locus to deleterious rearrangements. *Proc. Natl. Acad. Sci. USA* **87**: 1551-1555.

TISHKOFF, S. A., G. RUANO, J. R. KIDD and K. K. KIDD, 1996 Distribution and frequency of a polymorphic Alu insertion at the plasminogen activator locus in humans. *Hum. Genet.* **97**: 759-764.

WATKINS, W. S., C. E. RICKER, M. J. BANSHAD, M. L. CARROLL, S. V. NGUYEN *et al.*, 2001 Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. *Am. J. Hum. Genet.* 68: 738-752.

ZHANG, Y., K. M. DIPPLE, E. VILAIN, B. L. HUANG, G. FINLAYSON

et al., 2000 AluY insertion (IVS4-52ins316alu) in the glycerol kinase gene from an individual with benign glycerol kinase deficiency. *Hum. Mutat.* 15: 316-323.

Communicating editor: Y.-X. Fu